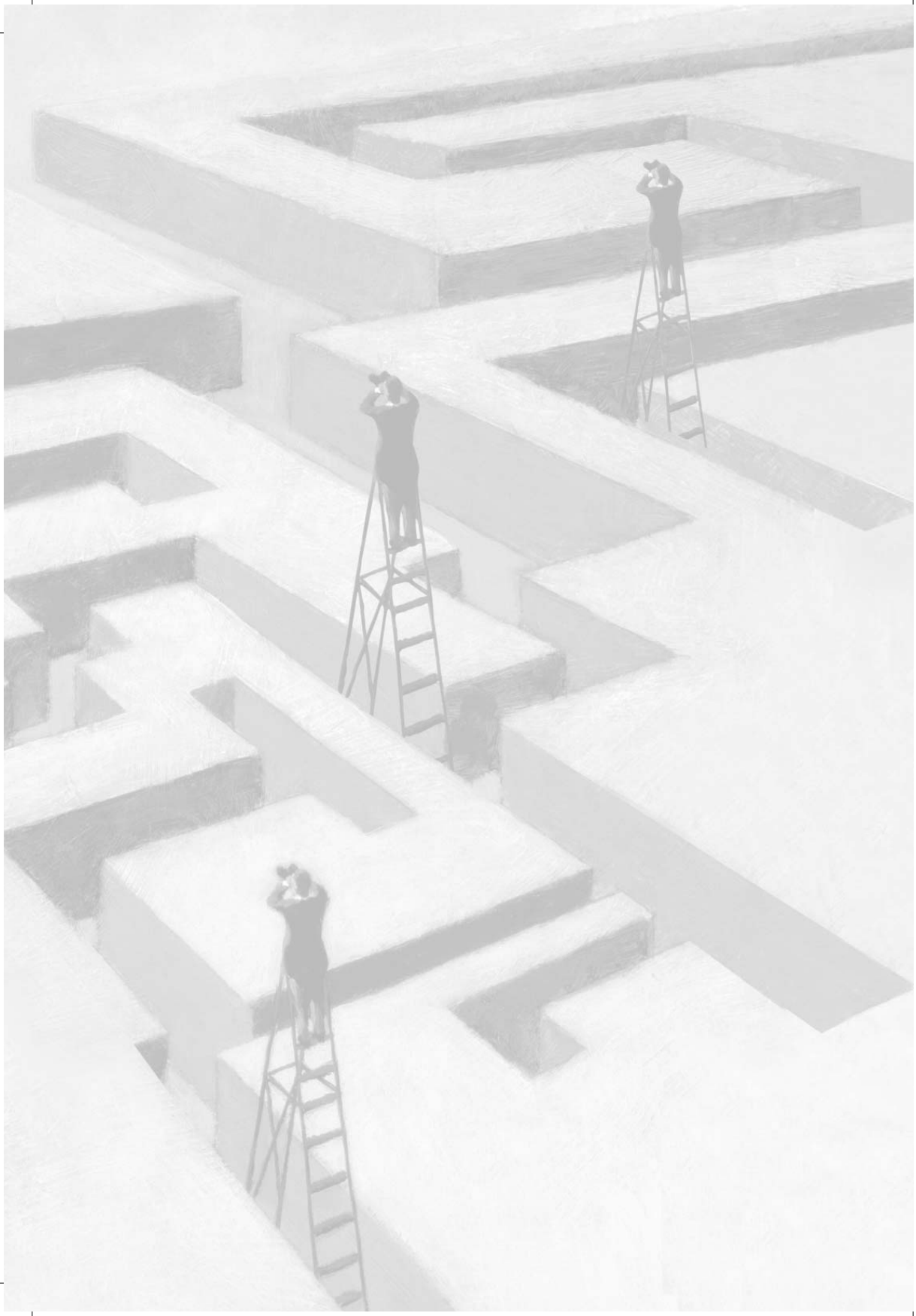
MIDCS – Um Método de Inferência Difusa para Classificação de Sonegadores Fiscais

3º Lugar

STRAUSS CUNHA CARVALHO*

* Graduado em Engenharia da Computação pelo Centro Universitário de Barra Mansa, Mestre em Engenharia Eletrônica e Computação pelo Instituto Tecnológico de Aeronáutica e Analista de Sistemas – Serpro SUPDE/Derjo.



MIDCS – Um Método de Inferência Difusa para Classificação de Sonegadores Fiscais

Resumo

A ideia de classificar os cidadãos com base em suas rendas e atribuir-lhes um tributo obrigatório, surgiu na Inglaterra, no fim do século XVIII, classificando-os manualmente. Atualmente, no Brasil, classificam-se os cidadãos por meio de sistemas computadorizados, armazenando as informações em Banco de Dados. Apesar dessa significativa evolução, o modo de classificação utilizado, até os dias de hoje, permanece ainda o mesmo, ou seja, fundamenta-se na bivalência da Lógica Clássica, incapaz de representar informações incertas e imprecisas. Assim, esse problema ocorre, também, na classificação de contribuintes fiscais da Receita Federal do Brasil (RFB), pois os algoritmos empregados se fundamentam na Lógica Clássica, ou seja, realizam a classificação por meio de uma nítida fronteira, classificando os contribuintes somente em duas categorias: não sonegadores e potenciais sonegadores, em malha fina. Esse fato faz com que os resultados, sob uma visão do mundo real, não satisfaçam as necessidades do usuário. Desse modo, o presente trabalho propõe um Método de Inferência Difusa para a Classificação de Sonegadores fiscais, denominado MIDCS, visando aumentar a eficiência no tratamento de incertezas e imprecisões nas recuperações e classificações de informações, a fim de tratá-las, qualitativamente,

de modo semelhante ao raciocínio humano. Assim, propiciando aos sistemas de Banco de Dados, por meio da utilização de termos qualitativos (linguísticos), as capacidades de recuperar, classificar e manipular informações, representando-as além das fronteiras da Lógica Clássica. O MIDCS utiliza-se de um Sistema de Inferência Difusa (SID), composto de uma base de regras e um mecanismo de inferência, aplicável, também, na tarefa de Classificação da etapa de Mineração de Dados (Data Mining – DM) do Processo de Descoberta de Conhecimento em Banco de Dados (Knowledge Discovery in Databases – KDD). O método proposto foi verificado em um estudo de caso envolvendo a classificação de 4.627.796 contribuintes do Imposto de Renda de Pessoas Físicas (IRPF), com base em uma réplica descaracterizada e desfragmentada da base de dados de desenvolvimento, a fim compará-lo com um método tradicional de consulta e classificação em Bancos de Dados. Portanto, após a realização dos experimentos, os resultados demonstraram que, diante da recuperação e classificação da informação, houve um contraste entre a aplicação do MIDCS, fundamentado na Lógica Difusa e no modelo convencional, com base na Lógica Clássica. Assim, propiciando uma classificação gradual entre os níveis de sonegação, não desprezando informações úteis à tributação e prevendo tendências de sonegação para posterior tomada de decisão, aumentando a precisão dos serviços do governo à sociedade.

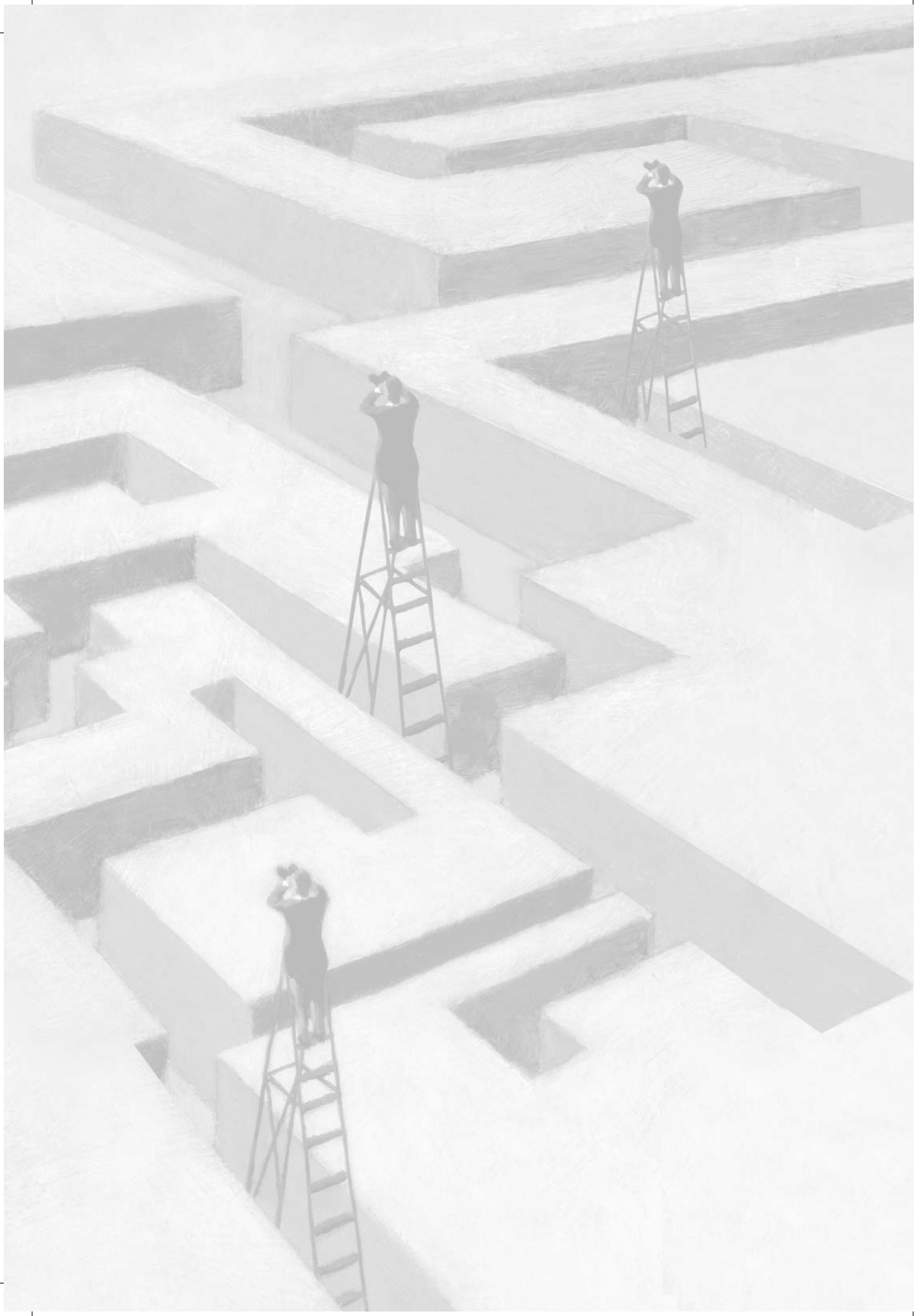
Palavras-Chave: Lógica Difusa. Banco de Dados. Inteligência Artificial. Classificação de Dados.

Lista de Figuras

Lista de Tabelas

Agradecimento

Ao Dr. Adilson Marques da Cunha, Professor Associado 4 da Divisão de Ciência da Computação do Instituto Tecnológico de Aeronáutica (ITA), pela confiança, paciência, e, principalmente, pelo ensinamento de seus conhecimentos e experiências que contribuíram, significativamente, para a minha formação acadêmica e pessoal.



MIDCS – Um Método de Inferência Difusa para Classificação de Sonegadores Fiscais

1 Introdução

A cada ano, a quantidade de dados armazenados nos Banco de Dados (BD) das organizações públicas ou privadas vem tendo um aumento diretamente proporcional às suas necessidades funcionais. Esses dados, intrinsecamente, têm um potencial para propiciar a melhoria dos processos empresariais e, consequentemente, aumentar a lucratividade. Assim, observa-se, nesses BD, valiosas informações guardadas, porém ocultas.

Além do grande volume de dados, as incertezas e as imprecisões, intrínsecas às informações do mundo real, também residem nos BD. Desse modo, ao recuperá-las, por meio dos métodos convencionais, obtêm-se informações inapropriadas e ou que não atendem, integralmente, às expectativas do usuário (PERES; BOSCARIOLI, 2002).

No entanto, segundo Ma (2006), os BD clássicos sofrem, frequentemente, com a incapacidade de armazenar, representar e manipular informações incertas e imprecisas. Isto se deve ao fato de que o armazenamento e a recuperação de informações fundamentam-se na Lógica Clássica (LC), fazendo com que os resultados das consultas,

sob uma visão do mundo real, não satisfaçam as necessidades dos tomadores de decisão.

Contudo, uma abordagem que vem se destacando nos últimos anos tem sido a utilização da Lógica Difusa (LD), em Inglês, *Fuzzy Logic* (FL), de modo a capacitar os BD a propiciarem a manipulação de informações incertas e imprecisas pelos seres humanos. Assim, permitindo que, diferentemente da LC, um elemento não seja classificado somente como falso ou como verdadeiro – bivalência.[1] Desse modo, a LD admite a classificação de informações sem uma fronteira definida, ou seja, propiciando um gradiente de valores entre o falso e o verdadeiro, assim, representando uma variação entre a completa falsidade e a verdade absoluta – multivalência.[2] Nesse contexto, o foco desse trabalho é investigar os métodos, as técnicas e as ferramentas que propiciam a aplicação da LD no âmbito dos Bancos de Dados.

Portanto, a presente pesquisa tem por objetivo investigar a aplicação da LD na recuperação e classificação de informações, propondo a concepção e a implementação de um Método de Inferência Difusa para classificação em BD, visando aumentar a sua eficiência no tratamento da incerteza e da imprecisão na recuperação e classificação de informações, a fim de tratá-las, qualitativamente, de modo semelhante ao raciocínio humano.

Assim, este trabalho compõe-se de cinco seções. Nesta primeira Seção introdutória, apresentam-se, resumidamente, o tema, o problema e o objetivo.

A Seção 2 desenvolve a fundamentação teórica que serve de base para o trabalho. Nela, são descritas, suscintamente, a tarefa de Classificação da Mineração de Dados e uma introdução da Lógica Difusa – técnica de Inteligência Artificial empregada no método proposto. Na Seção 3, segue uma investigação de três principais pesquisas relacionadas a esse trabalho.

1 A bivalência é a utilização de dois valores: algo é falso ou verdadeiro, zero ou um, portanto, exclui-se o meio termo (SIMÕES; SHAW, 2007).

2 Na multivalência, existe um espectro de opções entre o falso e o verdadeiro, portanto, considera-se o meio termo (SIMÕES; SHAW, 2007).

A aplicação do método proposto, por meio de um estudo de caso envolvendo 4.627.796 unidades experimentais, encontra-se descrita na Seção 4. Nela, aborda-se a classificação de sonegadores fiscais.

Na Seção 5, são os resultados obtidos com a aplicação do método proposto por meio de 34 experimentos. Por fim, a conclusão encontra-se na Seção 6.

2 Fundamentação teórica

Esta Seção aborda os principais conceitos teóricos que servem de base para a aplicação do método de inferência proposto. Nela, descreve-se o escopo desta pesquisa na área de BD, apresentando a Mineração de Dados e a sua tarefa de Classificação. Descreve-se, também, uma introdução da Lógica Difusa – técnica de Inteligência Artificial empregada no método proposto.

2.1 Mineração de dados

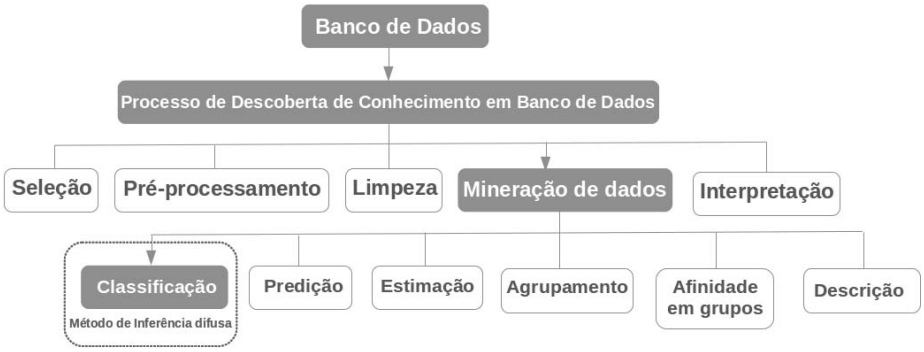
A Mineração de Dados (MD), em Inglês, Data Mining (DM), é uma etapa do Processo de Descoberta de Conhecimento em Banco de Dados ou, em inglês, Knowledge Discovery in Databases (KDD).[3] Para tal, ela utiliza-se de técnicas, entre as quais, se destacam: Árvores de Decisão, Regras de Associação, *Clustering*, Redes Neurais, Estatísticas e a técnica abordada neste trabalho – a Lógica Difusa, aplicada à tarefa de Classificação de Dados.

Berry e Linoff (2004) afirmam que a utilização da Lógica Difusa (LD), como técnica na etapa de MD, torna-se útil quando se deseja descobrir informações em um BD onde predominam dados numéricos que precisam ser classificados em conjuntos. E, de acordo com Galindo et al. (2001), a aplicação da LD na MD tem sido citada em alguns trabalhos, porém, de uma forma não explícita, quanto ao seu uso. Nesse contexto, por meio de uma representação gráfica, o diagrama da Figura 1 apresenta as restrições de escopo, na área de Banco de Dados, da presente

3 O processo de KDD tem por objetivo a extração do conhecimento implícito e previamente desconhecido e a busca da informação potencialmente útil nos dados, por meio da interseção de diferentes áreas (FAYYAD et al., 1996), (HAN; KAMBER, 2006).

pesquisa, convergindo até onde se aplica o método proposto, ou seja, na tarefa de classificação de dados. Assim, os blocos destacados nesta figura representam as subáreas de interesse do trabalho.

Figura 1 – Escopo do trabalho dentro da área dos Bancos de Dados (Autor)



Fonte: Elaboração do autor.

Embora recomendado à etapa de MD, o método proposto nesse trabalho pode ser aplicado isoladamente, ou seja, não há a dependência de uma estrutura de MD previamente definida.

2.2 Classificação de dados

De acordo com Harrison (1998), a tarefa de classificação, em MD, consiste na construção de um modelo que se aplica aos dados não classificados, visando categorizá-los em classes. Assim, examina-se um elemento, classificando-o de acordo com uma classe previamente definida.

No entanto, a classificação tradicional, originalmente fundamentada na Lógica Clássica (LC), qualifica os elementos seguindo a premissa de que eles pertencem integralmente a uma classe, ou seja, despreza-se o meio termo, porventura, um potencial elemento integrante. Dessa forma, Rosch (1978) apud Branco (2004), por meio de evidências experimentais, observa que, ao contrário do que a lógica clássica implica, o processo de formação de categorias é influenciado pelas características dos seres humanos envolvidos nesse processo de classificação.

A partir de tal observação, o presente trabalho investiga a aplicação da LD, com o objetivo de propiciar uma Classificação dos dados, fundamentando-se no pensamento humano e ao mundo real. Ou seja, sem limitar-se a um fronteira definida, atribuindo, aos elementos, um gradiente de valores entre o falso e o verdadeiro, representando, assim, uma variação entre a completa falsidade e a verdade absoluta.

2.3 Inteligência Artificial (IA)

De acordo com Arariboia (1989), a Inteligência Artificial (IA) é o ramo do conhecimento que trata, entre outros, do projeto e da construção de máquinas inteligentes, dotando-as de habilidades que caracterizam a Inteligência Natural (IN) – inerente aos seres humanos.

Russel e Norving (2013) apresentam, em duas dimensões e duas abordagens, isto é, quatro definições para o estudo da IA: 1) a dimensão do pensamento e raciocínio; 2) a dimensão do comportamento; 3) a abordagem centrada em torno dos seres humanos; e 4) a abordagem centrada em torno da racionalidade.[4]

Assim, ao longo dos anos, diversas técnicas de IA têm sido estudadas e aprimoradas, a fim de emular o raciocínio humano, entre elas, destacam-se: 1) Redes Neurais Artificiais; 2) Algoritmos Genéticos; 3) Redes Bayesianas; 4) Cadeias de Markov; 5) Lógica Paraconsistente; e 6) Lógica Difusa (LD).

Nessas técnicas, a LD propicia o tratamento, nos BD, da incerteza e da imprecisão, a fim de classificar informações sem uma fronteira nítida. Assim, o autor deste trabalho propõe um método de classificação com base na LD, por meio de um Sistema de Inferência Difusa, descrito na próxima Subseção.

4 Na racionalidade, observa-se que os seres humanos não são perfeitos e não possuem o mesmo grau de sucesso em uma determinada habilidade mental adquirida. Assim, considerando os indivíduos que conheçam todas as regras do jogo de xadrez, nem todos os detentores desse conhecimento tornam-se mestres no jogo (RUSSEL; NORVING 2004).

2.3.1 Sistema de Inferência Difusa (SID)

Os Sistemas de Inferência Difusa (SID), Segundo Cox (1995), assim como os tradicionais Sistemas Especialistas (SE)[5] de suporte à decisão, fundamentam-se no fluxo de entrada, processamento e saída. Entretanto, o SID difere-se em duas propriedades: 1) o uso da incerteza na representação das variáveis de entrada e de saída; e 2) o processamento fundamentado em regras linguísticas.

Desse modo, um SID propicia, por meio de variáveis e regras linguísticas, previamente projetadas e armazenadas em sua base de regras, realizar inferências aproximando a decisão computacional à decisão humana. Ou seja, obtendo conclusões considerando a incerteza e a imprecisão.

Nas três Subseções seguintes, descrevem-se, resumidamente, os componentes de um SID.

2.3.1.1 Componente 1 – Fuzificador

O Fuzificador, primeiro componente de um SID, transforma as variáveis de entrada, captadas do mundo real, em elementos que tendem a pertencer a um determinado conjunto difuso. Associando cada elemento a um número real na respectiva função de pertinência, assim, expressando-os como uma medida de imprecisão.

2.3.1.2 Componente 2 – Mecanismo de Inferência

O Mecanismo de inferência, núcleo de um SID, fundamenta-se na LD e seu objetivo é mapear os conjuntos difusos de entrada – referentes aos antecedentes, em conjuntos difusos de saídas – referentes aos consequentes, por meio da implicação das regras do tipo $p \Rightarrow q$, previamente definidas na base de regras.

A base de regras é composta por um conjunto de sentenças linguísticas do tipo *If – Then*, geralmente, fornecidas por especialistas na

5 Sistemas Especialistas são programas de computador, desenvolvidos para disponibilizar as habilidades de um profissional para um leigo, emulando o pensamento de um especialista numa determinada área de conhecimento (SILER, 2005).

área de domínio, ela corresponde, fundamentalmente, ao desempenho computacional de um SID. Desse modo, ao projetar a base de regras, como apresentado em 1, deve-se analisar o número de elementos do conjunto l, totalizado por n, ou seja, a quantidade de regras do SID. Portanto, seu valor é indiretamente proporcional ao desempenho computacional. Assim, uma grande quantidade de conjuntos difusos F_u^l e G_v^l podem causar uma explosão combinatorial de n. Apresenta-se, em 1, a fórmula geral para a criação de uma base de regras.

$$R^l : IF\ u_1\ is\ F_1^l\ AND\ u_2\ is\ F_2^l \ldots AND\ u_n\ is\ F_n^l\ THEN\ v\ is\ G^l \quad \textbf{(1)}$$

Em que:

$l = \{1,2,\ldots,r\}$ é o conjunto de regras. Sendo r = número total de regras;

$u = \{u_1, u_2, \ldots, u_n\}$ é o conjunto das variáveis linguísticas de entrada;

F_u^l representa os conjuntos difusos das variáveis linguísticas de entrada;

$v = \{v_1, v_2, \ldots, v_n\}$ é o conjunto das variáveis linguísticas de saída; e

G_v^l representa os conjuntos difusos da variável linguística de saída.

2.3.1.3 Componente 3 – Defuzificador

Nesse componente, opcional em alguns SID, transforma-se o conjunto difuso inferido em um único valor real de saída. Ou seja, obtendo um único valor numérico discreto que melhor representa os valores difusos inferidos da variável linguística de saída.

Para tal, utilizam-se métodos de defuzificação e a sua escolha depende da aplicação e da consequente interpretação do valor de saída (YAGER, 1996). Assim, no SID desse trabalho, emprega-se o método Centroide, também conhecido como Centro de Área, apresentado em 2.

$$N\mu_{out}(\mu_j) \quad \textbf{(2)}$$

Em que:

μ_{out} = a área de uma função de pertinência; e μ_j = a posição do centroide da função de pertinência individual.

Geralmente, é preciso um especialista do domínio de conhecimento para calibrar as curvas das funções de pertinência dos conjuntos difusos de saída. Nesse caso, alguns aspectos intuitivos, segundo Mendel (1995), devem ser considerados: 1) plausabilidade; 2) simplicidade computacional; e 3) continuidade.

2.3.2 Teoria dos Conjuntos Difusos (TCD)

Introduzida por Zadeh (1965), a Teoria dos Conjuntos Difusos (TCD), uma extensão da Teoria clássica dos conjuntos (TCC), propicia, por meio do grau de pertinência de seus elementos, a representação de conjuntos existentes no mundo real cujo os limites não são bem definidos. Assim, um conjunto difuso representa um agrupamento impreciso e indefinido, em que a transição de não pertinência para pertinência é gradual, não abrupta.

2.3.2.1 Números Difusos

Na TCC, a pertinência de um elemento x a um conjunto S é definida como:

$$x \in S \qquad \textbf{(3)}$$

Ou seja, expresso pela função características, em 4, dado um conjunto S em um universo U, os elementos desse universo possuem duas possibilidades – pertencem (verdadeiro) ou não pertencem (falso) ao subconjunto.

$$\mu_S(x) = \begin{cases} 0, & se \quad x \square S \\ 1, & se \quad x \in S \end{cases}$$

$$\mu_S(x) : U \rightarrow \{0,1\} \qquad \textbf{(4)}$$

Zadeh (1965) propôs, por meio da generalização da função características, uma ampla representação, permitindo-a assumir um número infinito de valores no intervalo de $[0,1]$. Dessa forma, em um conjunto universo U e um subconjunto difuso S, em que $S \subset U$, S define uma função de pertinência, demonstrada em 5, que associa a cada elemento $x \in U$, um grau $\mu_S(x)$ no intervalo $[0,1]$.

$$\mu_S(x): U \rightarrow [0,1] \quad \textbf{(5)}$$

Assim, um conjunto difuso S é uma coleção de pares $S=\{(x,\ \mu_s(x))\ \forall\ x \in U\}$, em que:

$\mu_S(x)$ é o grau de pertinência do elemento x com o conjunto S, ou seja, um elemento pode pertencer a mais de um conjunto difuso, com diferentes graus de pertinência.

2.3.3 Variáveis linguísticas

As variáveis linguísticas propiciam, por meio de uma descrição linguística qualitativa, uma caracterização de forma aproximada para fenômenos complexo e mal definidos – inerentes à comunicação humana. Assim, não se utilizando de variáveis quantificadas usualmente empregadas em termos matemáticos convencionais.

Formalmente, uma variável linguística caracteriza-se por uma $5-upla\ \ (N,\tau(N),U,G,M)$, em que:

N = nome da variável linguística;

$\tau(N)$ = o conjunto de valores que representam N, ou seja, os nomes dos valores linguísticos;

U = o universo de discurso;

G = a regra sintática que gera os valores de N, compondo os termos $\tau(N)$; e

M = a regra semântica que associa cada valor gerado por G a um conjunto difuso em N.

No estudo de caso apresentado na Seção 4, utilizam-se seis variáveis linguísticas: 1) renda; 2) despesas com educação; 3) despesas médicas; 4) número de dependentes; 5) omissão; e 6) sonegação. Desse modo, totalizando 25 conjuntos difusos.

2.3.4 Funções de pertinência

A função de pertinência determina o grau de possibilidade de um elemento pertencer a um determinado conjunto difuso, associando cada elemento $x \in U$ a um número real $\mu_S(x)$, no intervalo $[0,1]$.

Elas representam os conjuntos difusos em universos discretos ou contínuos, por meio dos valores que uma variável linguística pode assumir. Assim, funções de pertinência triangulares, trapezoidais e $\cos^2(x)$ são contínuas, porém com pontos não diferenciáveis. Funções gaussianas e sigmoides são contínuas e diferenciáveis em todos os pontos. Em razão de implicidade matemática, como apresentada na Equação 6 e, consequentemente, eficiência computacional, as funções triangulares e, também, trapezoidais, empregadas nesse trabalho, têm sido amplamente utilizadas. No entanto, uma vez compostas por segmentos de linha reta, não apresentam suavidade[6] em alguns pontos, culminando, posteriormente, resultados diferentes na inferência.

$$triangular(x;a,b,c)=\begin{cases} 0, & x \le a \\ \dfrac{x-a}{b-a}, & a \le x \le b \\ \dfrac{c-x}{c-b}, & b \le x \le c \\ 0, & c \le x \end{cases} \quad \textbf{(6)}$$

2.4 Lógica Difusa (LD)

O principal objetivo da Lógica Difusa (LD) é prover uma fundamentação teórica para lidar com proposições imprecisas, propiciando um raciocínio aproximado (ZADEH, 1965) apud (ROSS, 2004).Desse modo, a LD inspira-se nos conceitos da Lógica Clássica (LC), estendendo-a por meio da substituição das funções bivalentes por funções de pertinência. Com isso, uma proposição composta do tipo $se\ x$ é A então y é B tem a função de pertinência $\mu_{A \Rightarrow B}(x,y)$ que mede o grau de verdade da implicação entre $x\ e\ y$.

Ou seja, na LD, diz-se que dada uma proposição composta, sua veracidade ocorre se houver um grau de similaridade diferente de zero entre a proposição 1 e a proposição seguinte, resultando em uma consequente com um grau de similaridade não nulo. Dessa forma,

6 A classe das funções suaves corresponde àquelas funções que possuem derivadas de todas as ordens. Isto quer dizer que seu gráfico não apresenta quebras, saltos ou bicos.

LD propicia descrever o mundo em que se vive sem o dualismo do verdadeiro e do falso, permitindo valores intermediários entre os dois extremos.

Portanto, segundo Rabuske (1995), a LD viola o princípio da não contradição ao permitir proposições conflitantes variando de 0 a 1. E, ao permitir que proposições assumam infinitos de valores no intervalo de $[0;1]$, viola-se, também, o princípio do terceiro excluído (KLIR, 1997) apud (SOUZA, 2007).

2.4.1 Inferência

Ao se utilizar a LC como base, a LD propicia a realização de inferências por meio de regras de produção compostas por antecedentes e consequentes formadas por conjuntos difusos. Tais regras devem ser construídas, fundamentando-se no conhecimento de especialistas do domínio do problema a ser resolvido (ARTERO, 2009).

Embora haja diferentes estratégias de inferência, emprega-se, no SID aplicado ao estudo de caso desse trabalho, a implicação de Mamdani. Nessa estratégia de implicação, também denominada correlação de mínimo ou truncamento, correlaciona-se o consequente com a premissa verdadeira, truncando o grau de pertinência do consequente, como apresentado em 7.

$$\mu_{P\rightarrow Q}(x,y) = min\left\{\mu_P(x), \mu_Q(y)\right\} \quad \textbf{(7)}$$

3 Trabalhos relacionados

3.1 A Lógica Difusa e os Banco de Dados relacionais

Nos últimos anos, pesquisadores dos campos da Ciência da Computação têm investigado aplicações da Lógica Difusa (LD) em projetos de Banco de Dados (BD), motivados, principalmente, pela capacidade de armazenar, recuperar e manipular, por meio de termos linguísticos, informações incertas e imprecisas. Entretanto, constatou-se, nos trabalhos investigados e suas respectivas referências que não há um consenso sobre a aplicabilidade da LD na área de BD.

Portanto, de modo a reduzir o escopo da pesquisa, na Subseção a seguir, serão apresentados apenas os trabalhos que empregaram um Sistema de Inferência Difusa (SID) como classificador em BD fundamentar no modelo lógico relacional.[7]

3.1.1 Sistema de Inferência Difusa (SID) como classificador

A aplicação de um SID nos BD tem o objetivo, como apresentado na Seção 3.1, de aproximar a decisão computacional à decisão humana, por meio de regras linguísticas e procedimentos de inferência, a fim de obter conclusões. Desse modo, um SID, quando utilizado para a classificação de dados, necessita da construção do componente Mecanismo de inferência, que determina a forma de avaliação das regras; e de sua base de regras, que determina a estrutura do sistema (BRANCO, 2004).

Durante a realização dessa revisão bibliográfica, ao investigar as pesquisas relacionadas, o autor motivou-se por trabalhos que, do mesmo modo, adotaram um SID para a recuperação e classificação de informações em BD Relacionais.

O primeiro trabalho, de Branco (2004), aplicável na etapa de MD do processo de KDD, propõe um SID para problemas de classificação em geral. No entanto, mais adequado na área de *marketing* dirigido. Nesse primeiro trabalho, apresenta-se uma metodologia que, a partir de uma base de regras difusas obtidas por um algoritmo de aprendizado com a ajuda de um especialista, propicia a classificação por meio de consultas difusas. Desse modo, usando tabelas auxiliares, as consultas retornam uma lista ordenada dos registros que, segundo os conceitos difusos, possuem pertinência às respectivas classes.

Como exemplo, Branco (2004) apresenta, em 8, uma regra geral utilizada, atribuindo pesos às classes para a formação da base de regras.

$$IF\ X_i\ is\ A_{ik}\ THEN\ classe\ is\ \omega_j\ com\ peso\ \Phi_i(A_{ik},\omega_j). \quad \textbf{(8)}$$

O trabalho de Branco (2004) é importante no contexto desta pesquisa pois na implementação do Mecanismo de inferência de seu

7 Um modelo lógico é uma descrição de um BD no nível de abstração visto pelo usuário. Assim, o modelo lógico depende do tipo particular de SGBD utilizado. No modelo lógico relacional, os dados estão organizados na forma de tabelas (HEUSER, 2009).

SID: 1) utiliza o MATLAB – MatWorks documentation (2003), motivando modo de implementação do SID; e 2) descreve a aplicação de um SID na tarefa de classificação de etapa de MD.

O segundo trabalho investigado é o de Penteado (2009), no qual se propõe um método alternativo às consultas convencionais aos BD. Para tal, um SID utilizando-se do modelo Mamdani, implementado em Structured Query Language (SQL) padrão e armazenado em uma Visão (*View*)[8] no BD, permite a recuperação de informações segundo os preceitos da LD. No segundo trabalho, um SID composto por tabelas auxiliares armazena os valores linguísticos das variáveis de entrada e saída, de acordo com o domínio do problema. No estudo de caso, utilizaram-se, como variáveis de entrada para a classificação: 1) o PIB – Produto Interno Bruto; 2) o nível de escolaridade; 3) o total da população; e 4) a quantidade de empresas.

Os resultados do trabalho de Penteado (2009) motivou o autor dessa pesquisa a conceber um SID para classificação em BD. Entretanto, sem empregar tabelas auxiliares para o armazenamento dos valores linguísticos das variáveis de entrada e de saída.

No trabalho de Hudec e Vujosevic (2012), terceiro trabalho investigado na abordagem de SID como classificadores, apresenta-se uma metodologia composta por consultas e tarefas de classificação difusas. Nele, por meio de um estudo de caso envolvendo um BD climatológico das cidades da Slovênia, utilizam-se como variáveis de entrada para a classificação: 1) o comprimento das estradas *(road)*; e 2) o número de dias de neve *(snow)*.

Observa-se, no citado trabalho, a classificação de um município por meio de três conjuntos difusos da variável linguística de saída denominada *Maintenance*. Como exemplo, a cidade de Cebovce pertence à classe *(S – Small Maintenance)* com grau $0{,}267$ e à classe *(M – Medium Maintenance)* com grau $0{,}733$, resultando em um $P = 0{,}267 \times 0{,}1 + 0{,}733 \times 0{,}5 = 0{,}3932$. Assim, o passo de Classificação

8 Seja D um banco de dados e V uma visão sobre D, isto é, uma visão cuja definição X é uma função sobre D. Ou seja, trata-se de uma consulta dinâmica implementada em Linguagem SQL (DATE, 2004).

do MIDCS, descrito na Seção 4, fundamenta-se neste modo de atribuição de classes aos elementos.

Os trabalhos de pesquisa investigados mostram que, ao se usar os SID como classificadores de dados, observa-se a versatilidade da LD, na área da ciência da computação. Portanto, motivado pelas características de um SID e por não ser necessário investir esforços quanto à adaptação do BD, exceto quanto à criação de tabelas adicionais, decidiu-se por essa abordagem para conceber um método de inferência difusa para classificação em BD, a ser apresentado, por meio de um estudo de caso, na próxima Seção.

4. A aplicação do método

Esta seção apresenta, como principal contribuição dessa pesquisa, o Método de Inferência Difusa para Classificação de Sonegadores Fiscais, o MIDCS. Nela, mostra-se a aplicação do método proposto aplicado a um estudo de caso envolvendo a classificação de 4.627.796 contribuintes do imposto de renda de pessoas físicas, fundamentando-se em uma réplica descaracterizada de um Banco de Dados da RFB. De modo a garantir sua confidencialidade, os dados utilizados no estudo de caso não correspondem aos dados reais do ambiente de produção.

As próximas subseções descrevem o cenário do estudo de caso, a aplicação dos passos e subpassos do método proposto e, por fim, o desempenho computacional de seus algoritmos.

4.1 Cenário do Estudo de Caso

A ideia de classificar os cidadãos fundamentando-se em suas rendas e atribuir-lhes um tributo obrigatório surgiu na Inglaterra, no fim do século XVIII, quando, ameaçada por Napoleão Bonaparte, necessitava de recursos para o financiamento da guerra (RECEITA FEDERAL DO BRASIL, 2014).

Atualmente, no Brasil, esse tributo denomina-se imposto de renda e obriga o contribuinte, pessoa física ou pessoa jurídica, a deduzir, com base em suas informações financeiras, uma porcentagem de sua renda e lucros anuais para o governo federal. Desse modo, no ano de 2013,

26.034.621 contribuintes enviaram suas declarações de imposto de renda à Receita Federal do Brasil (RFB), até a data-limite 30/04/2013.

Após o recebimento das declarações de imposto de renda, a RFB iniciou um processamento computacional para analisar os 26.034.621 contribuintes, a fim de encontrar irregularidades nas informações coletadas e, consequentemente, potenciais sonegadores. Assim, classificaram-se os contribuintes por meio de duas classes: 1) não sonegadores; e 2) potenciais sonegadores, em malha fina. Isto ocorreu porque os algoritmos empregados se fundamentaram na Lógica Clássica, ou seja, realizaram a classificação por meio de uma nítida fronteira definida entre as duas categorias.

No entanto, nem todos os contribuintes pertencentes à classe 2, em malha fina, caracterizaram-se sonegadores fiscais. Nesse caso, o cidadão pode, antes de ser notificado pela RFB, corrigir os dados enviando uma declaração retificadora. Assim, após um novo processamento, o contribuinte pode, porventura, ser incluído na classe 1, não sonegadores. Dessa forma, após n processamentos, observou-se que, no mês de agosto de 2013, 1.320.000 contribuintes, (5,07%), faziam parte da classe 2. Posteriormente, no mês de dezembro de 2013, após o envio de declarações retificadoras, reduziu-se a classe 2 para 711.00 contribuintes, (2,73%), conforme a Tabela 1.

Tabela 1 – Resultado da classificação dos contribuintes em 2013

Classe 1	24.714.621	94,93%	não sonegadores	Agosto
Classe 2	1.320.000	5,07%	potenciais sonegadores (malha)	
Classe 1	26.033.910	97,27%	não sonegadores	Dezembro
Classe 2	711.000	2,73%	potenciais sonegadores (malha)	

Fonte: Elaboração do autor.

Nota-se que, por um período de tempo indeterminado, a RFB necessitou realizar n processamentos, a fim de analisar as declarações presentes na classe 2, naquele momento. Isto tornou-se um problema, pois, ao se examinar n vezes cada um dos contribuintes em malha fina, nesse alto volume de dados, houve uma alta demanda de processamento computacional e, em alguns casos, intervenção humana.

Um outro sintoma observado refere-se ao cruzamento dos dados coletados nas declarações com os dados oficiais, constantes nos Bancos de Dados da RFB. Assim, se o valor de um dado numérico coletado ultrapassar o seu limite preestabelecido, não existiu ainda um mecanismo para tratar os níveis de discrepância. Ou seja, o contribuinte teve a mesma classificação, se o valor foi próximo ou distante do limiar permitido.

Nesse cenário, percebeu-se uma evolução da classificação e tributação dos cidadãos, passando de classificação manual, utilizada na Inglaterra do século XVIII, para uma classificação e tributação informatizada, por meio dos computadores. No entanto, apesar dessa significativa evolução, o modo de classificação utilizado até 2013 permaneceu ainda o mesmo, ou seja, fundamentou-se na bivalência da Lógica Clássica (LC). Dessa forma, desprezando-se as informações incertas, imprecisas e intrínsecas do mundo real, tratadas pelo MIDCS, descrito a seguir.

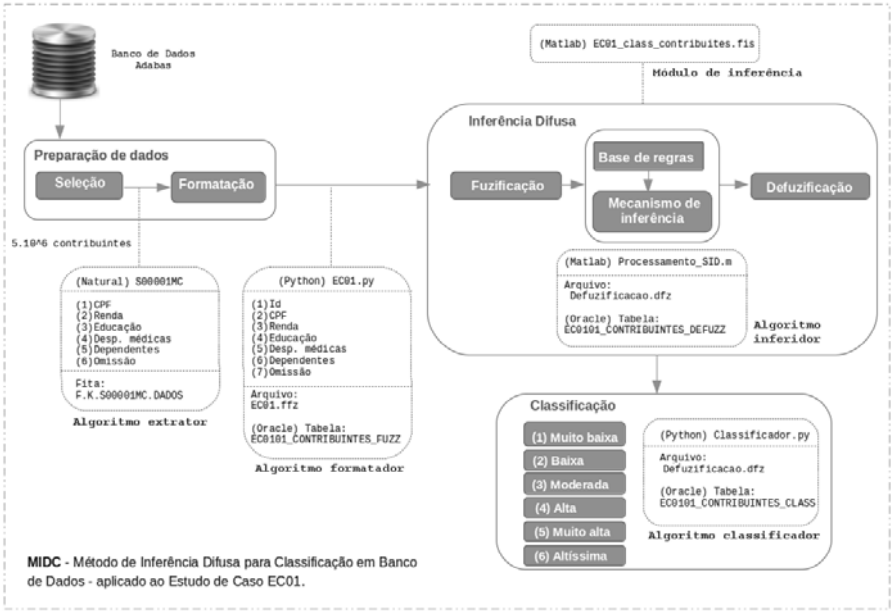
4.2 Passos do MIDCS

Com base no cenário apresentado, idealizou-se utilizá-lo como estudo de caso. Portanto, por meio do MIDCS, utilizando-se a Multivalência da LD, propõe-se uma classificação gradual em níveis de sonegação.

Assim, no ano de 2013, em um universo de 26.034.621 declarações enviadas à RFB, utilizando-se de um tratamento estatístico para garantir a confiabilidade das amostras, 2.943 unidades experimentais são suficientes. Contudo, além da confiabilidade das amostras, desejou-se verificar, também, o desempenho computacional do MIDCS. Portanto, empregaram-se, no estudo de caso, 4.627.796 amostras, ou seja, contribuintes do imposto de renda.

O diagrama da Figura 2 apresenta o MIDCS, aplicado ao estudo de caso, com seus passos, subpassos e algoritmos, descritos nas seções seguintes.

Figura 2 – Diagrama – Passos do MIDCS aplicado ao estudo de caso EC01 (Autor)



4.2.1 Passo 1 – Preparação de dados

O primeiro passo do MIDCS, composto por dois subpassos – Seleção e Formatação, tem por objetivo selecionar os dados relevantes à classificação e, posteriormente, formatá-los, de modo a corrigir erros e ou ruídos, propiciando, ao SID, no passo seguinte, a confiabilidade dos dados de entrada.

Assim, como apresentado na Figura 2, gerou-se um arquivo inicial, por meio do algoritmo extrator, o S00001mc, escrito em Linguagem Natural – Software *ag* documentation (2014). Ele foi executado somente em ambiente de desenvolvimento, selecionando os registros com os campos relevantes à classificação visando, de antemão, definir as variáveis linguísticas do MIDCS. Posteriormente, o algoritmo formator, EC01.py, escrito na Linguagem Python – Python documentation (2014),

formatou, eliminando caracteres especiais, os dados dos 4.627.796 de contribuintes presentes no arquivo.

Por fim, produziu-se o arquivo fuzificador, denominado EC01.ffz, composto dos seguintes campos formatados: 1) identificador do contribuinte; 2) CPF (descaracterizado); 3) divergência entre a renda declarada; 4) despesas com educação; 5) número de Dependentes; e 6) fator de omissão de informação. Adicionalmente, o conteúdo do arquivo foi gravado em uma tabela, em um BD Oracle – Oracle Database (2014).

4.2.2 Passo 2 – Inferência difusa

O segundo passo do MIDCS, dividido em três subpassos, compõe-se de um Sistema de Inferência Difusa (SID), descrito na subseção 3.1. Desse modo, apresentam-se, a seguir, os componentes do SID desenvolvido para este estudo de caso.

4.2.2.1 Subpasso 2.1 – Fuzificação

Iniciou-se este subpasso projetando os conjuntos difusos de entrada, de acordo com as variáveis linguísticas, previamente definidas no passo anterior. Simulando o conhecimento de um especialista em sonegação fiscal, projetou-se, para a classificação dos contribuintes, cinco variáveis linguísticas e 19 conjuntos difusos com os respectivos tipos de funções de pertinência, como apresentado na Tabela 2.

A partir das informações da Tabela 2, implementou-se, no MATLAB – MatWorks documentation (2003), por meio do *Fuzzy Logic Toolbox*: as variáveis linguísticas; os conjuntos difusos; e as funções de pertinência.

Tabela 2 – Variáveis linguísticas de entrada, conjuntos difusos e tipos de funções de pertinência empregadas no estudo de caso

Variável Linguística	Conjunto Difuso	Função de Pertinência
Renda	1-Baixa	Triangular
	2-Média	
	3-Alta	
	4-Muito alta	
	5-Altíssima	Trapezoidal
	6-Muito altíssima	
Despesas Educação	1-Baixa	Triangular
	2-Moderada	
	3-Alta	
Despesas Médicas	1-Baixa	Triangular
	2-Moderada	
	3-Alta	
	4-Muito alta	Trapezoidal
Dependentes	1-Pouco	Triangular
	2-Moderado	
	3-Alto	
Omissão	1-Pouca	Triangular
	2-Moderada	
	3-Alta	

Fonte: Elaboração do autor.

Posteriormente, projetou-se os conjuntos difusos de saída, necessários ao passo de Defuzificação. Para tal, usou-se apenas uma variável linguística de saída, denominada sonegação, composta por seis conjuntos difusos referentes a seis níveis de sonegação, como apresentado na Tabela 3.

Tabela 3 – Variável linguística de saída, conjuntos difusos e tipo de função de pertinência empregada no estudo de caso

Variável Linguística	Conjunto Difuso	Função de Pertinência
Sonegação	1-Muito baixa (ou nula)	Triangular
	2-Baixa	
	3-Moderada	
	4-Alta	
	5-Muito alta	
	6-Altíssima	

Fonte: Elaboração do autor.

4.2.2.2 Subpasso 2.2 – Base de regras

A base de regras compõe-se de um conjunto de sentenças linguísticas do tipo *If – Then*. Portanto, simulando, de modo simplificado, o conhecimento de um especialista em sonegação fiscal, foram criadas as regras, com base na fórmula geral, apresentada em (9).

$$R^l : IF\ u_1\ is\ F_1^l\ AND\ u_2\ is\ F_2^l\ AND\ u_3\ is\ F_3^l\ AND\ u_4\ is\ F_4^l\ AND\ u_5\ is\ F_5^l\ THEN\ v\ is\ G^l \quad \textbf{(9)}$$

Em que:

$l = \{1,2,3,... \ , n = 540\}$ é o conjunto de regras. Sendo que $n = 540$ corresponde ao total de regras, apresentado na Equação 10;

$u = \{u_1, u_2, u_3, u_4, u_5\}$ é o conjunto das cinco variáveis linguísticas de entrada;

F_u^l representa os 19 conjuntos difusos das variáveis linguísticas de entrada, apresentados na Tabela 2;

$v = \{v_1\}$ é o conjunto unitário da variável linguística de saída; e

G^l representa os seis conjuntos difusos da variável linguística de saída, apresentados na Tabela 3.

Multiplicando-se o total de variáveis linguísticas pelo total de conjuntos difusos que elas podem pertencer, obteve-se, para n, um total de 648 regras, como apresentado no cálculo da Equação (10). No entanto, de acordo com o conhecimento de especialistas e com informações de domínio público, utilizou-se um fator de ajuste k que representa a quantidade de regras irrelevantes à inferência. Desse modo, observaram-se 108 regras para esse fator, resultando um total de 540 regras.

$$\begin{aligned} n &= (F^l(u_1).F^l(u_2).F^l(u_3).F^l(u_4).F^l(u_5)) - k \\ n &= (6.3.4.3.3) - 108 \\ n &= 648 - 108 \\ n &= 540 \end{aligned} \quad \textbf{(10)}$$

Como exemplo, apresenta-se em (11), a regra R^{31}, que é uma das regras utilizadas no estudo de caso EC01. Nessa proposição, os termos linguísticos “baixa”, ’moderada’ e ’pouca’ atribuídos aos conjuntos difusos

"renda", "educação", "despesas médicas", "dependentes" e "omissão", resultaram em uma sonegação "baixa".

$$R^{31} = \begin{array}{r} IF\ renda\ is\ baixa\ AND\ educacao\ is\ moderada \\ AND\ despesas_{m}edicas\ is\ baixa\ AND\ dependente\ s\ is\ moderado \\ AND\ omissao\ is\ pouca\ THEN\ sonegacao\ is\ baixa \end{array} \quad \textbf{(11)}$$

4.2.2.3 Subpasso 2.3 – Mecanismo de inferência

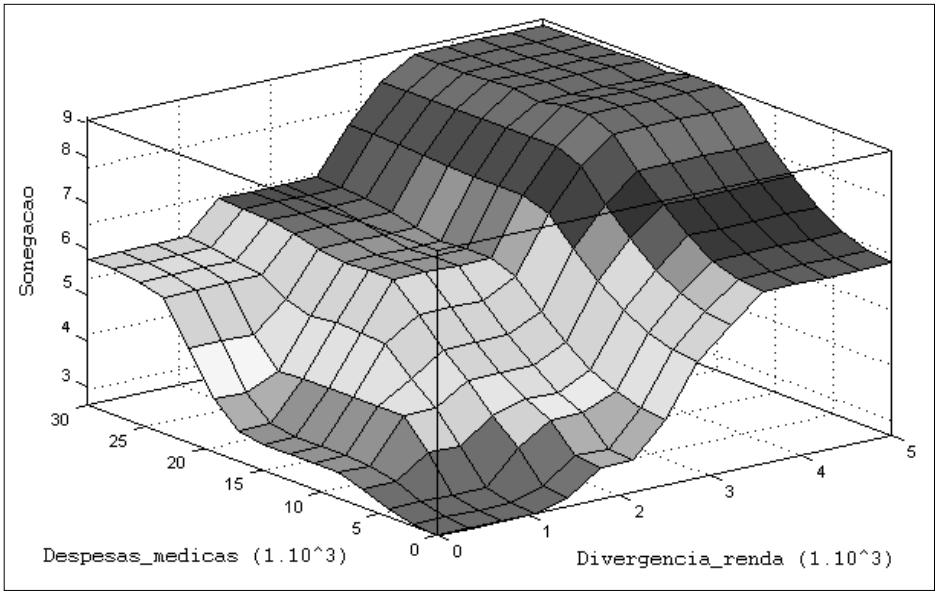
Nesse subpasso, implementa-se o mecanismo de inferência, componente núcleo de um SID, mapeando os conjuntos difusos de entrada, definidos no subpasso de fuzificação, em conjuntos difusos de saídas, definidos no passo de defuzificação, por meio da implicação das regras, previamente construídas no subpasso anterior.

Por meio da base de regras, previamente projetada, o mecanismo de inferência mapeou, para cada variável linguística, os conjuntos difusos de entrada, referentes aos antecedentes, apresentados na Tabela 2, em conjuntos difusos de saída, referentes aos consequentes, apresentados na Tabela 3.

Na Figura 3, apresenta-se um gráfico de superfície gerado no MATLAB – Matworks documentation (2003), por meio do *Fuzzy Logic Toolbox*. Nele, o eixo Z representa a variável linguística de saída, denominada sonegação.

Os eixos x e y representam, respectivamente, as variáveis linguísticas de entrada despesas médicas e divergência renda.

Figura 3 – Gráfico de superfície – despesas médicas x divergência renda x sonegação



Fonte: Elaboração do autor.

4.2.2.4 Subpasso 2.4 – Defuzificação

Nesse subpasso, necessitou-se calibrar as funções de pertinência, essencial para o resultado final da inferência. Para tal, utilizou-se o conhecimento de especialistas em sonegação fiscal e informações de domínio público sobre o assunto.

Ao fim do passo de inferência difusa, cada um dos 4.627.796 contribuintes foram submetidos ao SID para processamento. O algoritmo inferidor, denominado Processamento_SID.m, gerou o arquivo defuzificador EC01.dfz. Este arquivo é composto pelos CPFs descaracterizados de cada contribuinte e pelos respectivos valores numéricos correspondentes à função de pertinência da variável linguística sonegação, anteriormente inferida.

Por fim, os dados presentes no arquivo defuzificador foram inseridos em uma tabela, denominada ec0101_contribuintes_defuzz, em um BD Oracle, concluindo este passo do MIDCS.

4.2.3 Passo 3 – Classificação

Neste último passo do MIDCS, um algoritmo classificador, cuja complexidade de tempo é $O(n.m)$, utiliza os dados inferidos contidos no arquivo defuzificador (.dfz). Este arquivo, gerado no subpasso de defuzificação, possui os identificadores únicos de cada elemento e o seu respectivo resultado, ou seja, um valor numérico inferido pelo SID.

Na Figura 4, descreve-se o pseudocódigo do algoritmo classificador. Assim, ele se fundamenta nas funções de pertinência de saída, anteriormente definidas no subpasso de defuzificação, atribuindo, para cada elemento, em relação aos seus respectivos conjuntos difusos, um coeficiente de classificação, representado por c_x. Portanto, cada elemento pertenceu, de acordo com os preceitos difusos, a no mínimo 1 e no máximo 2 conjuntos difusos, representados por c_1 e por c_2, respectivamente.

Figura 4 – Pseudocódigo do algoritmo classificador do MIDCS (Autor)

```
   //Algoritmo classificador
 1 Atribua valores aos limites das F funções de pertinência de saída;
 2 Para as C classes (definidas pelas funções de pertinência de saída);
 3    Para os N elementos do arquivo  defuzificador (.ffz)
 4       Leia o 'id' do elemento e o resultado da inferência 'ri';
 5       Calcule a função triangular 'ft';
 6       Calcule o grau de classificação 'gc';
 7       Grave, no arquivo classificador (.clf), 'id', 'ri, 'gc, e 'c';
 8       Insira, na tabela tb_class_fuzz, 'id', 'ri, 'gc, e 'c';
 9    Fim-para;
10    Imprima a quantidade de elementos classificados de 'c';
11    Imprima a quantidade de elementos não classificados de 'c';
12    Imprima a quantidade de elementos lidos; (p10 + p11)
13 Fim-para;
```

Fonte: Elaboração do autor.

Por fim, o algoritmo propiciou o cálculo e gravação do resultado da classificação de cada contribuinte no arquivo classificador (.clf) e a inserção na Tabela de Classificação, denominada ec0101_contribuites_class, em um BD Oracle.

O resultado da classificação, para cada contribuinte, foi composto por: 1) identificador único do elemento; 2) resultado da inferência do SID; 3) coeficiente de classificação; e 4) classe do respectivo contribuinte. Dessa forma, finalizando os passos do MIDCS.

4.3 Desempenho computacional dos algoritmos

Apresenta-se, na Tabela 4, o resultado do desempenho computacional dos algoritmos empregados em todos os Passo do MIDCS, aplicados ao estudo de caso, no qual se representam:

n , os 4.627.796 contribuintes;

b , as 540 regras do SID empregado;

m , os 6 conjuntos difusos de saída; e

k , as 4.627.796 chamadas ao módulo de inferência.

Tabela 4 – Algoritmos empregados no MIDCS – E/S e complexidade

Algoritmo	Nome	Complexidade	Tempo Gasto
extrator	s00001mc	$O(n)$	2 min
formatador	ec01.py	$O(n)$	17 min
inferidor	processamento_SID.m	$O(n) + k$	302 min
módulo de inferência	ec01_class_contribuintes.fis	$O(n) + O(b)$	0
classificador	classificador.py	$O(n.m)$	16 min

Fonte: Elaboração do autor.

Portanto, o tempo aproximado de processamento verificado, desde o primeiro passo do MIDCS, a preparação de dados, até o passo final, a classificação, para os 4.627.796 contribuintes foi de 337 minutos, aproximadamente 5 horas e meia.

5 Principais resultados

Essa Seção aborda os experimentos e os principais resultados obtidos, a partir da aplicação do Método de Inferência Difusa para Classificação de Sonegadores Fiscais (MIDCS) no estudo de caso, descrito na Seção 4. Nela, apresenta-se uma análise comparativa da classificação por meio do MIDCS, fundamentado na Lógica Difusa (LD), com o modelo convencional, com base na Lógica Clássica (LC). Nesse trabalho, convencionou-se, simplificadamente, um indivíduo como não sonegador fiscal ao possuir, integralmente, as seguintes características: 1) divergência da renda declarada abaixo de R$ 120,01; 2) despesas com educação inferior a R$ 1.545.51; 3) gastos com despesas médicas

inferior a R$ 150,01; 4) no máximo, 2 dependentes; e 5) apenas 1 fator de omissão.

Nas seções seguintes, serão descritos os 34 experimentos realizados.

5.1 Grupo 1 de experimentos

No primeiro grupo de experimentos, utilizaram-se oito contribuintes, a fim de classificá-los, por meio de uma consulta SQL, como sonegadores ou não sonegadores. Desse modo, utilizando-se o fator classificação de dados, com dois tratamentos: 1) com base na LC, resultando no subconjunto S_c – sonegador, ou no subconjunto $S_c^{'}$ – não sonegador; e 2) após a aplicação do MIDCS, resultando no subconjunto S_d.

Por meio da LC, como apresentado na segunda linha da Tabela 5, na primeira subconsulta SQL, selecionam-se as características de um indivíduo não sonegador fiscal. A segunda subconsulta verifica se o contribuinte possui indícios de sonegação, ou seja, pertence ao subconjunto $S_c^{'}$, não sonegador, assim, resultando em uma resposta binária – sim ou não.

Por meio do MIDCS, com base na LD, classifica-se um contribuinte por seu coeficiente de classificação em relação às n classes de sonegação que ele tende a pertencer, como apresentado na terceira linha da Tabela 5.

Tabela 5 – Grupo 1 de Experimentos – lógica clássica *versus* aplicação do MIDCS

	instrução SQL	**resultado**
lógica clássica	select fctb_cpf from (select * from ec0101_ contribuintes_fuzz where fctb_renda <= 0.12 and fctb_educacao <= (3.091/2) and fctb_ medicina <= 0.15 and fctb_dependentes <= 2 and fctb_omissao <= 1) where fctb_cpf = *contribuinte*;	sim ou não
após o MIDCS	select cctb_classe, cctb_grau_pertinencia, cctb_resultado_defuzz from ec0101_contribuintes _class where cctb_cpf = *contribuinte*;	classe(s) coef. classif.

Fonte: Elaboração do autor.

A seguir, descrevem-se os oito experimentos realizados. Na Tabela 6, são apresentados os seus resultados sumarizados, classificando-se os contribuintes por meio da Lógica Clássica, confrontando-a com a aplicação do MIDCS. Posteriormente, descreve-se uma análise dos experimentos.

Experimento 1: Observou-se que o contribuinte 12073192 classificou-se, por meio da LC, como não sonegador, pertencente ao subconjunto $S_c^{'}$. Após a aplicação do MIDCS, com o valor da inferência $0{,}39388$, ele pertenceu à classe $c1$ com coeficiente de classificação $c_x = 97{,}5051$, portanto, resultando em uma sonegação muito baixa ou nula.

Experimento 2: O contribuinte 527558233 classificou-se, por meio da LC, como sonegador, não pertencente ao subconjunto $S_c^{'}$. No entanto, após a aplicação do MIDCS, com o valor da inferência $0{,}39493$, ele pertenceu à classe $c1$ com coeficiente de classificação $c_x = 97{,}2329$, portanto, resultando em uma sonegação muito baixa ou nula.

Tabela 6 – Resultados do grupo 1 de experimentos – classificação lógica clássica *versus* aplicação do MIDCS

Exp.	Contribuinte	Lógica Clássica	após MIDCS
1	12073192	não sonegador	inferência difusa - 0, 39388 classe c1 - 97, 5051 sonegação muito baixa ou nula
2	527558233	sonegador	inferência difusa - 0, 39493 classe c1 - 97, 2399 sonegação muito baixa ou nula
3	255975488	sonegador	inferência difusa - 5, 85571 classe c4 - 93, 7943 classe c5 - 12, 3784 sonegação alta sonegação muito alta
4	626324748	sonegador	inferência difusa - 2, 73944 classe c2 - 57, 3463 classe c3 - 36, 5394 sonegação baixa sonegação moderada
5	129467834	sonegador	inferência difusa - 1, 9461 classe c2 - 96, 2954 sonegação baixa
6	4931916986	não sonegador	inferência difusa - 0, 4990 classe c1 - 70, 9596 sonegação muito baixa ou nula
7	730521989	sonegador	inferência difusa - 3, 77657 classe c3 - 95, 804 sonegação moderada
8	624982347	sonegador	inferência difusa - 1, 58941 classe c2 - 68, 1209 sonegação baixa

Fonte: Elaboração do autor.

Experimento 3: O contribuinte 255975488 classificou-se, por meio da LC, como sonegador, não pertencente ao subconjunto $S_c^{'}$. Após a aplicação do MIDCS, com o valor da inferência $5{,}85571$, ele pertenceu às classes $c4$ e $c5$ com coeficientes de classificação $93{,}7943$ e $12{,}3784$, respectivamente, portanto, resultando em uma sonegação alta com tendência de $12{,}37$ à muito alta.

Experimento 4: O contribuinte 626324748 classificou-se, por meio da LC, como sonegador, não pertencente ao subconjunto $S_c^{'}$. Após a aplicação do MIDCS, com o valor da inferência $2{,}73944$, ele pertenceu às classes $c2$ e $c3$ com coeficientes de classificação $57{,}3463$ e $36{,}5394$, respectivamente, portanto, resultando em uma sonegação baixa com tendência de $36{,}53$ à moderada.

Experimento 5: O contribuinte 129467834 classificou-se, por meio da LC, como sonegador, não pertencente ao subconjunto $S_c^{'}$. Após a aplicação do MIDCS, com o valor da inferência $1{,}9461$, ele pertenceu à classe $c2$ com coeficiente de classificação $96{,}2954$, portanto, resultando em uma sonegação baixa.

Experimento 6: O contribuinte 4931916986 classificou-se, por meio da LC, como não sonegador, pertencente ao subconjunto $S_c^{'}$. No entanto, após a aplicação do MIDCS, com o valor da inferência $0{,}4990$, ele pertenceu à classe $c1$ com coeficiente de classificação $70{,}9596$, portanto, resultando em uma sonegação muito baixa, porém, não nula.

Experimento 7: O contribuinte 730521989 classificou-se, por meio da LC, como sonegador, não pertencente ao subconjunto $S_c^{'}$. Após a aplicação do MIDCS, com o valor da inferência $3{,}77657$, ele pertenceu à classe $c3$ com coeficiente de classificação $95{,}8004$, portanto, resultando em uma sonegação moderada.

Experimento 8: O contribuinte 624982347 classificou-se, por meio da LC, como não sonegador, não pertencente ao subconjunto $S_c^{'}$. No entanto, após a aplicação do MIDCS, com o valor da inferência $1{,}58941$, ele pertenceu à classe $c2$ com coeficiente de classificação $68{,}1209$, portanto, resultando em uma sonegação baixa.

5.1.1 Análise dos resultados do grupo 1 de experimentos

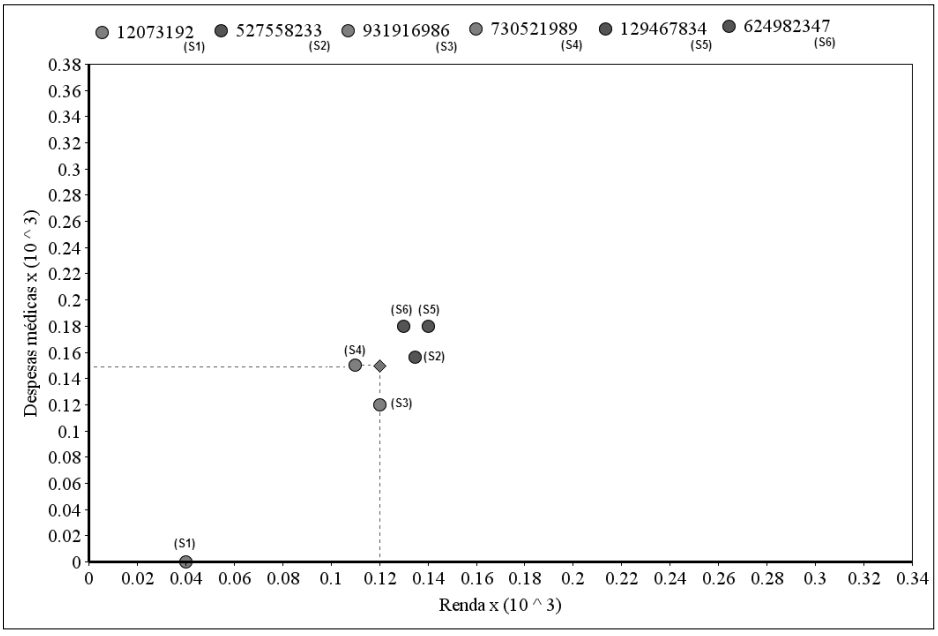
Observaram-se nos oito experimentos realizados que, por meio da LC, as consultas classificaram os contribuintes admitindo somente duas possibilidades: 1) verdadeiro, para o subconjunto $S_c^{'}$, não sonegador; e 2) falso, para o subconjunto S_c, sonegador. Assim, tal classificação binária dos contribuintes não foi capaz de prever a tendência de quão

sonegador ou não sonegador eles são, desprezando informações úteis à tributação e posterior tomada de decisão.

Portanto, 2 contribuintes classificaram-se como sonegadores, pertencentes ao subconjunto S_c e 2 como não sonegadores, pertencente ao subconjunto S_c'.

No gráfico das Figura 5, apresentam-se os resultados da classificação de seis contribuintes, por meio da LD, ou seja, delimitados por nítidas fronteiras. O gráfico despreza os dados de omissão dos contribuintes pois foram igual a 0. Assim, apresentam-se, nos eixos x e y, respectivamente, os valores das Rendas e os valores das despesas médicas dos seis contribuintes.

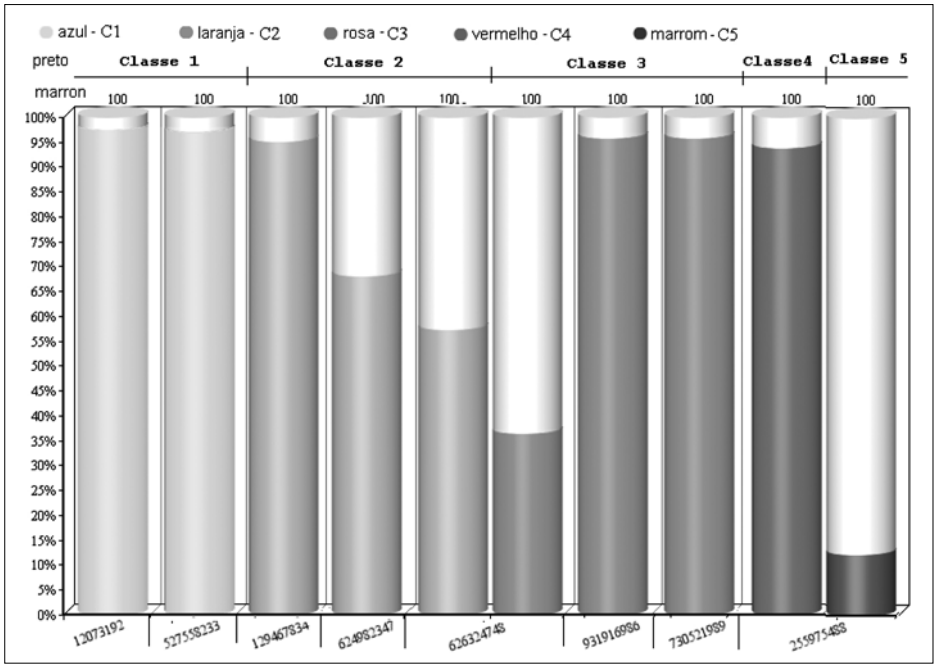
Figura 5 – Classificação de seis contribuintes por meio da LC (Autor)



Fonte: Elaboração do autor.

No gráfico da Figura 6, apresentam-se os resultados dos experimentos realizados perante a aplicação do MIDCS, classificando gradualmente oito contribuintes, por meio de seus respectivos coeficientes de classificação.

Figura 6 – Classificação de oito contribuintes após a aplicação MIDCS (Autor)



Fonte: Elaboração do autor.

Observou-se que, mediante o coeficiente de classificação, seis contribuintes pertenceram, respectivamente, a uma única classe de sonegação. Assim, tomando como exemplo o contribuinte 12073192, ele classificou-se perante a classe $c1$ – sonegação muito baixa ou nula, com coeficiente de classificação igual a 97,5051, ou seja, com altíssima pertinência à classe.

Entretanto, dois contribuintes pertenceram, simultaneamente, a duas classes de sonegação. Assim, o contribuinte 626324748 classificou-se nas classes $c2$ – sonegação baixa e $c3$ – sonegação moderada, com coeficiente de classificação 57,3463 e 36,5394, respectivamente. Ou seja, com maior pertinência à classe $c2$, porém, com moderada inclinação à classe $c3$.

O contribuinte 255975488 classificou-se perante às classes $c4$ – sonegação alta e $c5$ – sonegação muito alta, com coeficiente de classificação 93,7943 e 12,3784, respectivamente. Ou seja, com grade pertinência à classe $c4$, porém, com baixa tendência à classe $c5$.

Portanto, os resultados do grupo 1 de experimentos de classificação, realizados após a aplicação do MIDCS, propiciaria aos tomadores de decisão realizarem: 1) uma tarifação gradual e, consequentemente, com maior precisão ao nível de sonegação; 2) uma avaliação das tendências de sonegação dos contribuintes; 3) uma premiação, de algum modo, aos contribuintes classificados com coeficiente nulo de sonegação; e 4) uma penalização, de algum modo, aos contribuintes classificados com grau elevado de sonegação.

5.2 Grupo 2 de experimentos

O segundo grupo de experimento visou classificar os 4.627.796 contribuintes totalizando-os em sonegadores e em não sonegadores. Para isso, foram utilizados os fatores: 1) classificação de dados; e 2) classes (c). O primeiro Fator se utilizou de 2 Tratamentos com base na LC. O segundo fator se utilizou de seis tratamentos, com base na aplicação do MIDCS.

Por meio da LC, foram usadas duas consultas que totalizaram a quantidade de contribuintes não sonegadores, representados pelo subconjunto $(S_c^{'})$ e sonegadores, representado pelo subconjunto (S_c).

Após a aplicação do MIDCS, totalizam-se os contribuintes, representados pelo subconjunto (S_d), fundamentando-se nas seis classes de sonegação c1, c2, c3, c4, c5 e c6.

Experimento 9: Este experimento utilizou-se de 4.627.796 unidades experimentais cujos resultados e as respectivas consultas SQL se encontram sumarizados na Tabela 7.

Tabela 7 – Experimento 4 – Resultado dos experimentos

	instrução	Resultado
lógica clássica	select count(fctb_cpf) as total_sonegacao_nula from ec0101_contribuintes_fuzz where fctb_renda <= 0.12 and fctb_educacao <= (3.091/2) and fctb_medicina <= 0.15 and fctb_dependentes <= 2 and fctb_omissao <= 1;	3.725.304
	select count(fctb_cpf) as total_sonegacao from ec0101_contribuintes_fuzz where fctb_renda > 0.12 and fctb_educacao > (3.091/2) and fctb_medicina > 0.15 and fctb_dependentes > 2 and fctb_omissao > 1;	902.492
após o MIDCS	select count(cctb_cpf) as total_classe_c1 from ec0101_contribuintes_class where cctb_classe = 'c1';	3.898.823
	select count(cctb_cpf) as total_classe_c2 from ec0101_contribuintes_class where cctb_classe = 'c2';	564.890
	select count(cctb_cpf) as total_classe_c3 from ec0101_contribuintes_class where cctb_classe = 'c3';	391.177
	select count(cctb_cpf) as total_classe_c4 from ec0101_contribuintes_class where cctb_classe = 'c4';	213.995
	select count(cctb_cpf) as total_classe_c5 from ec0101_contribuintes_class where cctb_classe = 'c5';	135.449
	select count(cctb_cpf) as total_classe_c6 from ec0101_contribuintes_class where cctb_classe = 'c6';	5

Fonte: Elaboração do autor.

5.2.1 Análise dos resultados do grupo 2 de experimentos

Neste nono experimento, observou-se que, por meio da LC, a classificação binária, mais uma vez, resultou em uma nítida fronteira entre os subconjuntos $S_c^{'}$ de não sonegadores, totalizando 902.492 contribuintes, e o subconjunto S_c de sonegadores, totalizando 3.725.304 contribuintes. Entretanto, por meio da aplicação do MIDCS, observou-se que o subconjunto S_d resultante propiciou uma classificação gradual entre as seis classes de sonegação. O resultado da totalização, para cada uma das seis classes de sonegação, representa a quantidade de contribuintes que possuem o coeficiente de classificação c_x, variando de 0 a 100, às respectivas classes. Desse modo, tomando-se como exemplo a classe $c1$, afirma-se que houve 3.898.823 contribuintes com c_x variando de 0 a 100, ou seja, resultando todos os contribuintes que pertencem à classe $c1$, independentemente do coeficiente de classificação c_x. O mesmo modo de classificação ocorreu para as outras cinco classes.

5.3 Grupo 3 de experimentos

O terceiro grupo de experimentos, com base no subconjunto S_d resultante, após a aplicação do MIDCS, visou classificar os 4.627.796 contribuintes utilizando, como fator, o coeficiente de classificação c_x, por meio de seis tratamentos fundamentados nas seis classes de sonegação. Desse modo, realizaram-se 25 experimentos que consistiram em ajustar o coeficiente de classificação c_x, por meio do fator de ajuste fa de 0% a 100%, com base em seus respectivos conjuntos difusos. Assim, totalizando os n_x contribuintes que pertencem à respectiva classe de sonegação e possuam o coeficiente de classificação maior ou igual que c_x, ou seja, $c_x >= fa$.

A seguir, descrevem-se os 25 experimentos.

Experimentos 10 a 24: Os resultados de 15 experimentos, realizados por seis tratamentos, representados pelas classes de sonegação, encontram-se sumarizados na Tabela 8. As linhas dessa tabela representam 15 coeficientes de classificação, ou seja: c_x = (10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 75, 80, 85 e 95). As colunas representam o total de contribuintes classificados n_x, fundamentando-se nas seis classes de sonegação C = (c1, c2, c3, c4, c5 e c6).

Tabela 8 – Resultado dos experimentos obtidos em seis classes ajustando-as em 15 coeficientes de classificação (c_x)

Exp	c_x	$n_x(C1)$	$n_x(C2)$	$n_x(C3)$	$n_x(C4)$	$n_x(C5)$	$n_x(C6)$
Exp 10	$>= 95$	3.862.339	32.677	73.645	7.782	0	0
Exp 11	$>= 85$	3.867.490	151.805	118.606	51.411	0	0
Exp 12	$>= 80$	3.867.750	200.881	130.668	57.338	0	0
Exp 13	$>= 75$	3.868.035	240.057	142.017	63.587	0	0
Exp 14	$>= 70$	3.877.779	275.586	154.256	70.154	0	0
Exp 15	$>= 60$	3.878.467	343.009	177.033	83.650	37.412	2
Exp 16	$>= 50$	3.878.740	392.448	220.318	99.845	44.685	4
Exp 17	$>= 45$	3.878.808	417.376	248.249	108.944	47.059	5
Exp 18	$>= 40$	3.878.898	435.788	262.131	124.755	49.801	5
Exp 19	$>= 35$	3.888.189	465.889	276.296	156.702	52.987	5
Exp 20	$>= 30$	3.888.470	481.783	292.363	164.349	62.899	5
Exp 21	$>= 25$	3.888.779	497.700	304.344	171.540	68.886	5
Exp 22	$>= 20$	3.888.804	513.149	316.305	180.124	76.294	5
Exp 23	$>= 15$	3.889.079	528.602	328.081	188.504	84.788	5
Exp 24	$>= 10$	3.889.170	535.420	339.607	196.342	93.136	5

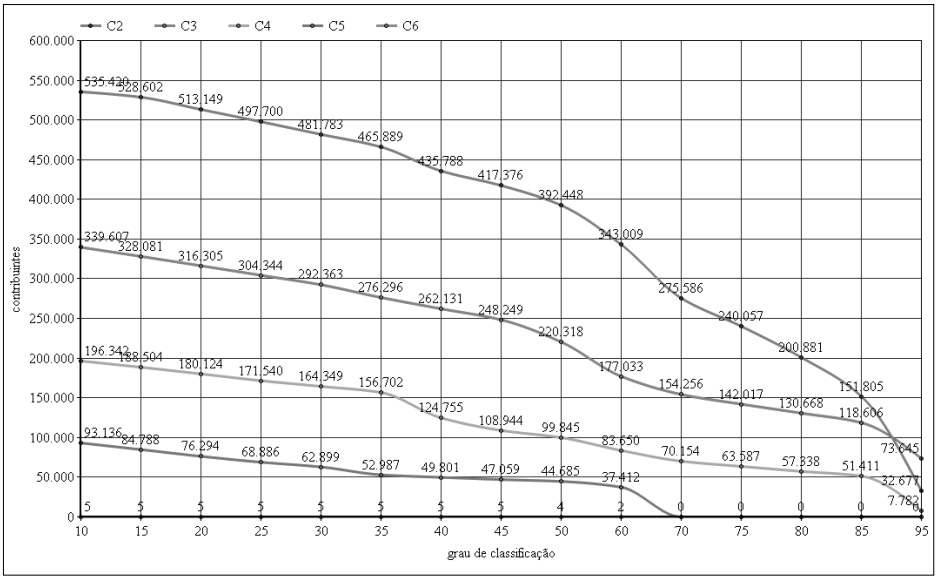
Fonte: Elaboração do autor.

Verificou-se, nos experimentos, que o ajuste do coeficiente de classificação c_x propiciou uma classificação gradual dos contribuintes em relação às classes de sonegação. Ou seja, proporcionalmente, quanto maior o valor de c_x, maior é a pertinência do contribuinte à classe.

Portanto, como exemplo, no experimento 10, ao se utilizar um $c_x >= 95$, resultou em um total n_x de contribuintes que pertenceram integralmente a cada classe. Do mesmo modo, no experimento 24, ao se utilizar um $c_x >= 10$, resultou no total de contribuintes classificados a partir de uma sutil inclinação de pertencer à respectiva classe.

O gráfico da Figura 7 apresenta os resultados dos 15 experimentos.

Figura 7 – Gráfico de classificação dos contribuintes ajustando c_x em relação às classe c2, c3 e c4 (Autor)



Fonte: Elaboração do autor.

Experimentos 25 a 34: Os resultados de 10 experimentos, realizados por meio de seis tratamentos, representados pelas classes de sonegação, encontram-se sumarizados na Tabela 9. As linhas dessa tabela representam 10 coeficientes de classificação, ou seja: c_x = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10). As colunas representam o total de contribuintes classificados n_x, com base nas seis classes de sonegação C = (c1, c2, c3, c4, c5 e c6).

Porém, ao contrário dos experimentos anteriores, classificaram-se os contribuintes com o coeficiente de classificação menor ou igual a c_x. Desse modo, representando a tendência do mesmo pertencer às respectivas classes.

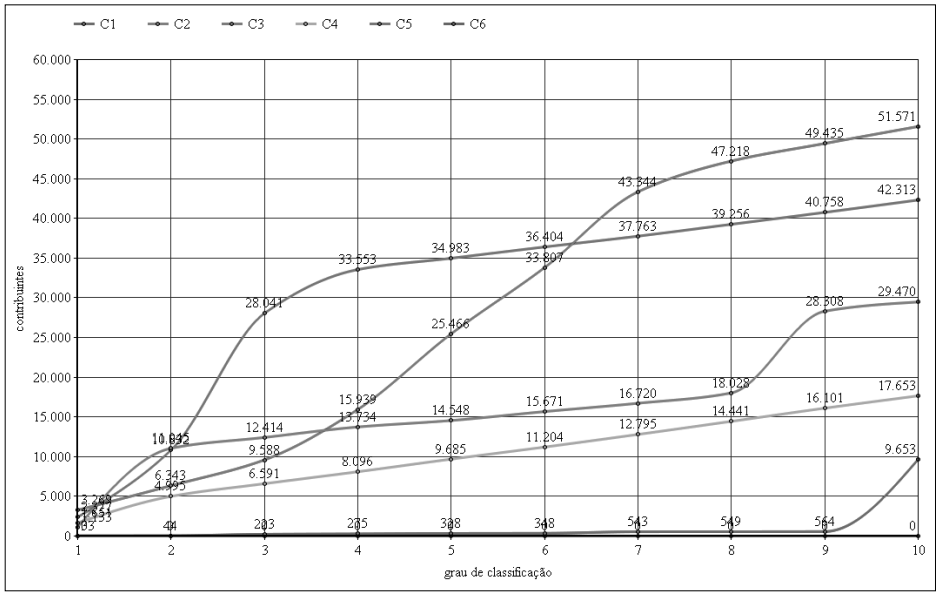
Tabela 9 – Resultado dos experimentos obtidos em seis classes ajustando-as em 15 coeficientes de classificação (c_x)

Exp	c_x	$n_x(C1)$	$n_x(C2)$	$n_x(C3)$	$n_x(C4)$	$n_x(C5)$	$n_x(C6)$
Exp 25	<= 10	9.653	29.470	51.571	17.653	42.313	0
Exp 26	<= 9	564	28.308	49.435	16.101	40.758	0
Exp 27	<= 8	549	18.028	47.218	14.441	39.256	0
Exp 28	<= 7	543	16.720	43.344	12.795	37.763	0
Exp 29	<= 6	348	15.671	33.807	11.204	36.404	0
Exp 30	<= 5	328	14.548	25.466	9.685	34.983	0
Exp 31	<= 4	275	13.734	15.939	8.096	33.553	0
Exp 32	<= 3	223	12.414	9.588	6.591	28.041	0
Exp 33	<= 2	44	11.045	6.343	4.995	10.832	0
Exp 34	<= 1	33	1.133	3.269	1.651	2.377	0

Fonte: Elaboração do autor.

O gráfico da Figura 8 representa, mediante os 10 experimentos, o total de contribuintes classificados de acordo com suas tendências a pertencer à respectiva classe.

Figura 8 – Gráfico de classificação dos contribuintes ajustando C em relação às classes c1, c2, c3, c4, c5 e c6 (Autor)



Fonte: Elaboração do autor.

5.3.1 Análise dos resultados do grupo 3 de experimentos

Observou-se, nos 25 experimentos, que o ajuste do coeficiente de classificação c_x permitiu classificar, proporcionalmente, os n_x elementos que pertencem às respectivas classe de sonegação. Ou seja, o valor de ajuste c_x implicou a quantidade total de elementos dos conjuntos.

Portanto, com base no estudo de caso, para as classes $c1$ e $c2$, adotando-se uma classificação rigorosa dos contribuintes, utilizar-se-iam valores de c_x próximo de 100, classificando-os com maior pertinência às classes. Entretanto, ao se adotar uma classificação branda nessas duas classes, ajustar-se-ia o valor de c_x próximo de 10, a fim de classificar os contribuintes com menor pertinência às classes.

6 Conclusão

Este trabalho de pesquisa teve como principal objetivo investigar e conceber um Método de Inferência Difusa para classificação de Sonegadores Fiscais, visando aumentar a sua eficiência no tratamento da incerteza e da imprecisão na recuperação e classificação de informações, a fim de tratá-las, qualitativamente, de modo semelhante ao raciocínio humano. Realizou-se uma revisão bibliográfica introdutória sobre Mineração de Dados e Lógica Difusa – Técnica de Inteligência Artificial utilizada para a tarefa de classificação no método proposto. Com o objetivo de complementar a revisão bibliográfica e demonstrar a aplicabilidade da fundamentação teórica, foram investigados os trabalhos de pesquisas relacionados e/ou relevantes para os objetivos desse trabalho.

A partir dos conceitos estudados, concebeu-se o Método de Inferência Difusa para Classificação de Sonegadores Fiscais (MIDCS), composto por três passos e sete subpassos. Ele possui um Sistema de Inferência Difusa (SID), com a finalidade de propiciar uma classificação de informações, usando termos qualitativos ou linguísticos, assim, recuperando dos Banco de Dados, informações incertas e imprecisas.

Para propiciar a verificação e a validação do método proposto, um estudo de caso, composto por 4.627.796 unidades experimentais, permitiu a aplicação do MIDCS em uma réplica descaracterizada do Banco de

Dados da Receita Federal do Brasil (RFB) a fim de classificar, sob os conceitos difusos, os contribuintes em seis diferentes níveis graduais de sonegação. Desse modo, foram feitos 34 experimentos, a fim de realizar uma análise comparativa do MIDCS, com base na Lógica Difusa, com o modelo convencional de recuperação de informação, com base na Lógica Clássica.

No grupo 1 de experimentos, observou-se que, ao contrário da classificação binária da Lógica Clássica, em que se dividiu os contribuintes em apenas duas classes distintas, sonegadores e não sonegadores, a aplicação do MIDCS propiciou uma classificação gradual, na qual se admitiu uma variação entre os seis níveis de sonegação. Dessa forma, pôde-se, além de classificar o contribuinte em uma classe, identificar a sua possível tendência a pertencer às classes adjacentes.

Assim, propiciando à classificação de contribuintes fiscais: 1) uma tarifação gradual e com maior precisão quanto ao nível de sonegação; 2) uma avaliação das tendências de sonegação; 3) uma premiação, quando classificados com coeficiente nulo de sonegação; e 4) uma penalização, quando classificados com grau elevado de sonegação.

No grupo 2 de experimentos, após a aplicação do MIDCS, totalizou-se, para cada uma das seis classes de sonegação, uma quantidade de contribuintes que pertenceu às respectivas classes com o coeficiente de classificação c_x, variando de 0 a 100. Nesse grupo 2, observou-se, na classe 1 – *Sonegação Muito Baixa ou Nula*, um aumento de 3,78% contribuintes classificados, representando 173.519 possíveis falsos-negativos.

Nesse grupo de experimentos, notou-se que, por meio do método proposto, houve um aumento de 173.519 contribuintes classificados com algum indício de não sonegação que poderiam ser analisados, utilizando-se o MIDCS com Lógica Difusa. Desse modo, reduzindo recursos envolvidos, incluindo tempo de reprocessamento e de análises por intervenção humana.

No grupo 3 de experimentos, percebeu-se que o ajuste do coeficiente de classificação, c_x, implicou, em cada uma das seis classes de sonegação, na quantidade total de contribuintes classificados. Assim,

permitiu-se, eventualmente, adotar uma classificação mais rigorosa dos contribuintes, ajustando-se valores de c_x mais próximo de 100, ou seja, compondo os contribuintes com maior pertinência às classes. Contudo, permitiu-se, também, uma classificação branda, ajustando o valor de c_x mais próximo de 10, ou seja, compondo os contribuintes com menor pertinência às classes.

Portanto, os resultados dos experimentos demonstraram que, diante da recuperação e classificação da informação, houve um contraste entre a aplicação do MIDCS e o modelo convencional. Assim, diferentemente da classificação binária, o MIDCS não se limita a nítida fronteira definida, admitindo uma variação gradual entre os níveis de sonegação, não desprezando informações úteis para a tributação e posterior tomada de decisão, aumentando a precisão dos serviços do governo perante a sociedade.

Referências

ARARIBÓIA, G. **Inteligência Artificial**: Um curso prático. 1. ed. São Paulo: Livros Técnicos e Científicos, 1989. 282 p. ISBN 9788521605911.

ARTERO, A. O. **Inteligência Artificial**: teoria e prática. 1. ed. São Paulo: Editora Livraria da Física, 2009. ISBN 978-85-7961-029-6.

BERRY, M.; LINOFF, G. **Data Mining Techniques**: For marketing, sales, and customer relationship management. Indianapolis: Wiley, 2004.

BRANCO, A. **Geração de Fuzzy Queries para Mineração de Dados**. 2004. 79 p. Tese (Doutorado em Engenharia Civil) – Universidade Federal do Rio de Janeiro, RJ.

COX, A. **Fuzzy Systems Handbook.** 1. ed. Boston: Academic Press Inc, 1995. DATE,

C. J. **Introdução a Sistemas de Banco de Dados**. 8. ed. São Paulo: Elsevier Editora, 2004. 896 p. ISBN 85-352-1273-6.

FAYYAD, U. et al. **Advances in Knowledge Discovery and Data Mining**. Menlo Park, Calif: AAAI Press, 1996. ISBN 9780262560979.

GALINDO, J. et al. **Relaxing the universal quantifier of the division in fuzzy relational databases**. n. 6, p. 713-742, 2001.

HARRISON, T. **Intranet data warehouse**. 1. ed. São Paulo: Bekerley Brasil, 1998.

HEUSER, C. A. **Projeto de Banco de Dados**. 1. ed. S o Paulo: Bookman, 2009. ISBN 979-85-7780-382-8.

HUDEC, M.; VUJOSEVIC, M. **Selection and classification of statistical data using fuzzy logic**. Conference on new techniques and technologies for statistics, Brussels, Proceedings...Brussels: INFOSTAT, p. 186–195, 2012.

KLIR, G. **Fuzzy Sets Theory**: Foundation and applications. Upper Saddle River, N. J: Prentice Hall, 1997. ISBN 9780133410587.

MA, Z. **Fuzzy Database Modeling of Imprecise and Uncertain Engineering Information**. 1. ed. Berlin: Springer, 2006. (Studies in Fuzziness and Soft Computing). ISBN 9783540306757.

MATHWORKS documentation. **Fuzzy Logic Toolbox**: user guide. USA, 2003. Disponível em: <http://www.mathworks.com/help/fuzzy/index.html>. Acesso em: 13 jun. 2014.

MENDEL, J. M. **Uncertain Rule-Based Fuzzy Logic System**: Introduction and new directions. Michigan: Prentice Hall, 1955. ISBN 9780130409690

ORACLE documentation. **Oracle Database**. USA, 2014. Disponível em: <http://docs.oracle.com/cd/E17781 01/index.htm>. Acesso em: 16 jun. 2014.

PENTEADO, F. B. L. **Método de Filtragem Fuzzy Para Avaliação de Bases de Dados Relacionais**. 2009. 100 p. Dissertação (Mestrado em Engenharia Elétrica) Escola de Engenharia de São Carlos – Universidade de São Paulo, SP.

PERES, S.; BOSCARIOLI, C. **Sistemas gerenciadores de banco de dados fuzzy**: Uma aplicação em recuperação de informação. Acta Scientiarum, v. 24, n. 6, p. 1733 -1743, 2002.

PYTHON documentation. **Python Programming Language**. USA, 2014. Disponível em: <https://docs.python.org/2/>. Acesso em: 13 jun. 2014.

RABUSKE, R. **Inteligência Artificial**. Santa Catarina: Universidade Federal de Santa e Catarina, 1995. 240 p. (Série didática). ISBN 9788532800251.

RECEITA FEDERAL DO BRASIL. **Memória Receita Federal**. Brasil, 2014. Disponível em: <http://www.receita.fazenda.gov.br/Memoria/irpf/historia/histPriomordios Mundo-.asp>. Acesso em: 15 jul. 2014.

ROSCH, E. **Cognition and Categorization**. 1. ed. Hillsdale, New. Jersey: Erlbaum, 1978. ROSS, T. J. Fuzzy Logic with Engineering Applications. 1. ed. New York: McGraw-Hill, 2004. ISBN 0-470-86074-X.

RUSSEL, S.; NORVING, P. **Inteligência Artificial**. 3 ed. Rio de Janeiro: Campus, 2013. 1021 p. ISBN 9788535211771.

SILER, W.; BUCKLEY, J. **Fuzzy expert systems and fuzzy reasoning**. 1. ed. Hoboken, N.J: Wiley-Interscience, 2005. ISBN 0-471-38859-9.

SIMOES, M. G.; SHAW, I. S. **Controle e modelagem fuzzy**. 9. ed. São Paulo: Blucher, 2007.

SOFTWARE ag documentation. **Natural and Adabas Database**. Germany, 2014. Disponível em: <https://empower.softwareag.com/Products/>. Acesso em: 13 jun. 2014.

SOUZA, C. **Teoria de conjuntos fuzzy e regressão logística na tomada de decisão para realização de cintilografia das paratiróides**. 2007. Dissertação (Mestrado em Saúde Pública), Universidade de São Paulo, SP.

ZADEH, L. A. **Fuzzy sets**. Information and Control, University of California, Berkeley, California, p. 338-353, 1965.