



APLICAÇÃO DE SISTEMAS DE INFERÊNCIA NEBULOSA COM O FUZZYMORPHIC.PL

patv

WAGNER ARBEX*, FABRÍZZIO CONDÉ DE OLIVEIRA†,
LUÍS ALFREDO VIDAL DE CARVALHO‡

**Empresa Brasileira de Pesquisa Agropecuária
R. Eugênio do Nascimento, 610, CEP 36038-330, Juiz de Fora, MG, Brasil*

†*Universidade Salgado de Oliveira
Av. dos Andradas, 731, CEP 36036-000, Juiz de Fora, MG, Brasil*

‡*Universidade Federal do Rio de Janeiro
Centro de Tecnologia, Bloco H-319, CEP 21945-970, Rio de Janeiro, RJ, Brasil*

Emails: arbex@cnpgl.embrapa.br, fabrizzioconde@gmail.com, alfredo@cos.ufrj.br

Abstract— Problems dealing with imprecise or uncertain features, e. g., problems of decision making, can be designed as fuzzy systems, since these systems allow the processing of subjective and qualitative argument, usually intrinsic in such problems. The fuzzyMorphic.pl is a tool for developing of fuzzy inference systems that enable the specification of the modeling and the implementation of these systems that enable the model description and implementation of fuzzy inference systems on different problems. In this specific case, the fuzzyMorphic.pl was used to develop and implement an application of fuzzy inference for decision making in a bioinformatics problem, the investigation of single nucleotide polymorphisms.

Keywords— Fuzzy inference system, Modeling of fuzzy system, Fuzzy inference, Implementation of fuzzy system, fuzzyMorphic.pl

Resumo— Problemas que tratam características imprecisas ou incertas, como, p. ex., problemas de tomada de decisão, podem ser modelados como sistemas nebulosos, uma vez que esses sistemas permitem o tratamento do raciocínio subjetivo e qualitativo, geralmente, intrínseco em problemas dessa natureza. O fuzzyMorphic.pl é uma ferramenta para desenvolvimento de sistemas de inferência nebulosa (SINs) que permite a descrição do modelo e a implementação de SINs em diferentes problemas. No caso em questão, o fuzzyMorphic.pl foi utilizado para desenvolver e implementar um aplicação de inferência nebulosa para a tomada de decisão em um problema bioinformática, a investigação de polimorfismos de base única.

Palavras-chave— Sistema de inferência nebulosa, Modelagem de sistema nebuloso, Inferência nebulosa, Implementação de sistema nebuloso, fuzzyMorphic.pl

1 Introdução

Sistemas de inferência nebulosa (SINs) permitem, p. ex., a implementação de modelos computacionais de mineração de dados para descoberta de conhecimento em bases de dados, com capacidade de processar informações imprecisas e qualitativas sendo, portanto, adequados em situações que exigem tomadas de decisão (de Almeida and Evsukoff, 2005).

Sob esse aspecto, o objetivo desse texto é o de apresentar um modelo computacional para tomada de decisão, desenvolvido e implementado com o fuzzyMorphic.pl (Arbex, 2009a; Arbex, 2009b), aplicado à investigação de polimorfismos de base única (*single nucleotide polymorphisms - SNPs*), em seqüências expressas de cDNA.

O modelo computacional desenvolvido utiliza-se de lógica nebulosa como base para a implementação de um sistema de inferência, auxiliar à tomada de decisão, que baseia-se em resultados prévios, obtidos por diferentes ferramentas de descoberta de SNPs e que apresentam resultados possivelmente conflitantes. Assim, o modelo é aplicado para auxiliar na tomada de decisão, no caso em que as informações sejam divergentes e, também,

na confirmação de informações coincidentes.

2 Sistemas de inferência nebulosa

SINs são modelagens do raciocínio subjetivo expressos em algoritmos computacionalmente executáveis por meio de regras de inferência, que, por sua vez, são provenientes e podem ser construídas a partir da teoria dos conjuntos nebulosos e da lógica nebulosa.

2.1 Fundamentos de conjuntos nebulosos e lógica nebulosa

O texto *Outline of a new approach to the analysis of complex systems and decision processes* (Zadeh, 1973) é considerado o marco inicial da lógica nebulosa, porém, diversos autores consideram que a lógica nebulosa iniciou-se em um trabalho anterior, *Fuzzy sets* (Zadeh, 1965), que definiu a teoria dos conjuntos nebulosos e, portanto, possibilitou a definição de uma lógica sobre esses conjuntos.

A subjetividade no raciocínio em geral, utilizada no cotidiano, sendo transmitida e perfeitamente compreendida entre interlocutores, é ex-

pressa em “termos e variáveis lingüísticas” (Zadeh, 1973) e não é expressa sob o ponto de vista da lógica clássica ou de qualquer abordagem matemática tradicional.

P. ex., o uso de adjetivos comuns que representam imprecisão ou incerteza, tais como, *alto*, *baixo*, *muito*, *pouco*, ou, ainda, relações e agrupamentos, como, *o conjunto das pessoas altas*, não podem ser expressos por essas abordagens, a menos que seja definido, com exatidão, o conceito ou o valor que, nesse caso, determine a altura, a partir da qual, uma pessoa pode ser considerada alta. Entretanto, para o padrão comum de entendimento entre dois interlocutores, não há nenhuma dificuldade de discernimento do que, p. ex., venha a ser uma pessoa alta. Na prática, os termos e variáveis lingüísticas aumentam a complexidade de um sistema de computação frente à capacidade de os sistemas de computação trabalharem com números ou outros valores exatos, discretos e, muitas vezes, excludentes.

Sob essa ótica, o “Princípio da Incompatibilidade” (Zadeh, 1973), retrata a incompatibilidade entre precisão e significância, a partir do aumento da complexidade do que se quer modelar:

com o crescimento da complexidade do sistema, a habilidade para descrever seu comportamento com precisão e significância (relevância) diminui até o ponto a partir do qual precisão e significância tornam-se características quase excludentes.

Decorre desse princípio a idéia de que trabalhar com valores incertos possibilita a modelagem de sistemas complexos, mesmo que, se for o caso, reduza a precisão do resultado, o que não retira a credibilidade. Se, por um lado, as incertezas, quando consideradas isoladamente, são indesejáveis, quando associadas a outras características dos sistemas, em geral, permitem a redução da complexidade do sistema e aumentam a credibilidade dos resultados obtidos (Klir and Yuan, 1995).

As abordagens clássicas para, p. ex., classificação ou agrupamento de elementos, são falhas quando os atributos avaliados se aproximam dos valores limites, representados por valores exatos e discretos. Isto é, supondo que sejam estabelecidos os limites $1,60\ m$ e $1,75\ m$ e, então, um indivíduo com estatura **inferior** a $1,60\ m$ será considerado *baixo*, um indivíduo com estatura **superior** a $1,75\ m$ será considerado *alto* e, por fim, serão considerados de estatura *mediana* os indivíduos cuja estatura esteja entre os limites estabelecidos. Assim, um indivíduo com a estatura de $1,74\ m$, deve ser considerado mediano e classificado juntamente com algum outro indivíduo que, por ventura, tenha a estatura de $1,61\ m$.

Essa decisão, lógica e matematicamente precisa, não parece razoável no contexto geral, devido

à subjetividade que a envolve e, de fato, não é, visto que o valor de $1,74\ m$ está muito mais próximo do limite, e do conjunto, determinado por $1,75\ m$, do que do limite de $1,60\ m$, ao qual foi associado. Todavia, como já foi dito, as abordagens clássicas da lógica e da matemática não possuem as ferramentas necessárias para tratar limites próximos, imprecisão ou incerteza.

A abordagem nebulosa propõe que sejam estabelecidas “intensidades” aos dados, quando comparados aos limites estabelecidos, o que, segundo a teoria dos conjuntos nebulosos e a lógica nebulosa, equivale a estabelecer um “grau de pertinência” à informação - objeto, valor, variável - variando entre zero e 1 em relação aos conjuntos aos quais a informação possa pertencer.

Como exemplificado na Tabela 1, o indivíduo A, com estatura de $1,74\ m$, poderia ser associado a um “grau de pertinência” de 0,95 com relação ao conjunto das pessoas altas, de 0,05 em relação ao conjunto das pessoas com estatura mediana e, ainda, igual a zero para o conjunto das pessoas baixas. Enquanto, para o indivíduo B, com estatura de $1,61\ m$, poderia ser associado um “grau de pertinência” igual a zero com relação ao conjunto das pessoas altas, de 0,05 em relação ao conjunto das pessoas com estatura mediana e 0,95 para o conjunto das pessoas baixas.

Tabela 1: Graus de pertinência de dois indivíduos em relação a estatura, para os conjuntos de estaturas alta, mediana e baixa.

Indivíduo	Altura	Estatura alta	Estatura mediana	Estatura baixa
A	1,74	0,95	0,05	0,0
B	1,61	0,0	0,05	0,95

Esse exemplo foi utilizado para aplicar a teoria dos conjuntos nebulosos e a proposta da lógica nebulosa, isto é, “interpretar” um certo dado de um indivíduo ou objeto, frente a um cenário com alguma incerteza, ou, ainda, algum grau de incerteza. O formalismo que permite essa aplicação se baseia na definição dos conjuntos nebulosos, que diz (Zadeh, 1965):

Seja X um espaço de pontos (objetos), com um elemento genérico de X denotado por x , então $X = \{x\}$. Um *conjunto nebuloso*, uma classe, A em X é caracterizado pela *função de pertinência*, função característica, $f_A(x)$ a qual associa a cada ponto em X um número real, no intervalo de $[0, 1]$, com esse valor, ou seja, o valor de $f_A(x)$ em x , representando o “grau de pertinência” de x em A . Assim, quanto mais próximo de 1 for o valor de $f_A(x)$, maior será o grau de pertinência de x em A .

Foi feita uma breve revisão sobre conjuntos e lógica nebulosos, visando o uso desses conceitos

em SINS e mediante a aplicação de tais conceitos em um problema, aparentemente, bastante simples. Todavia, devido a essa percepção, algumas questões devem ser observadas:

1. o raciocínio utilizado no exemplo de classificação das alturas dos indivíduos, apesar de simples, pode ser extrapolado para inúmeros problemas;
2. na verdade, é um exemplo simples, porém não tanto quanto parece. Casos como esses, se fossem triviais, as ferramentas clássicas poderiam facilmente resolvê-los, mas, quando se aproximam do raciocínio subjetivo para a interpretação e a extração de uma resposta, uma decisão, tornam-se bastante complexos e a aparente simplicidade é conferida pela modelagem por lógica nebulosa e por seu embasamento na teoria dos conjuntos nebulosos.

2.2 Sistemas de inferência nebulosa como apoio à decisão

A subjetividade intrínseca ao raciocínio é capaz de lidar com situações complexas, baseadas em informações imprecisas, incertas ou aproximadas e, para tanto, a estratégia adotada é a de utilizar “operadores humanos” também de natureza imprecisa, que são expressos por termos ou variáveis lingüísticas.

Essa perspectiva, essencialmente humana, de descrever ou tratar problemas, em geral, não permite uma solução em termos de números exatos, mas, por exemplo, conduz a solução a uma classificação, agrupamento ou agregação qualitativa em categorias ou possíveis conjuntos de soluções (de Almeida and Evsukoff, 2005). Essas soluções podem ser vistas como decorrência do “Princípio da Incompatibilidade” (Zadeh, 1973), já enunciado em 2.1.

A teoria dos conjuntos nebulosos e a lógica nebulosa são adequadas para representar, em termos matemáticos, a informação imprecisa, que pode ser expressa por um conjunto de regras lingüísticas e, caso exista a possibilidade de que os operadores humanos sejam organizados como um conjunto de regras da forma

se ANTECEDENTE então CONSEQÜENTE

logo, o raciocínio subjetivo pode ser construído em um algoritmo computacionalmente executável (Tanscheit, 2007) com capacidade de classificar, de modo impreciso, as variáveis que participam dos termos antecedentes e conseqüentes das regras, em conceitos qualitativos, e não quantitativos, o que representa a idéia de variável lingüística (de Almeida and Evsukoff, 2005). Assim, como sistemas capazes de processar de forma eficiente informações imprecisas e qualitativas, os modelos de inferência nebulosa são adequados em situações

que exigem tomadas de decisão (de Almeida and Evsukoff, 2005).

A arquitetura de um SIN, em geral, apresenta três grandes etapas: a conversão dos dados precisos de entrada em valores nebulosos (*fuzzyfication*), a inferência e a conversão dos resultados de saída, de valores nebulosos para grandezas numéricas precisas (*defuzzyfication*). A Figura 1 traz um exemplo dessa arquitetura (Cox, 1994; de Almeida and Evsukoff, 2005; de Carvalho, 2005; Goldschmidt and Passos, 2005; McNeill and Thro, 1994; Tanscheit, 2007).

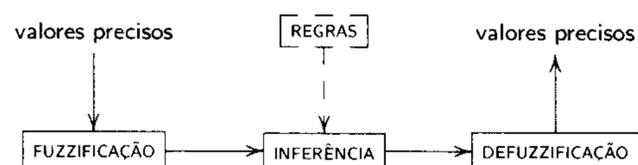


Figura 1: Estrutura generalizada de um SIN.

3 Inferência nebulosa para a descoberta de SNPs em seqüências expressas de cDNA com o fuzzyMorphic.pl

Os projetos de seqüenciamento de genomas revelaram mais variações e maior complexidade do que inicialmente previa-se. Uma das variações e particularidades dos genomas são os SNPs, isto é, pares de bases, em uma única posição no DNA genômico, que se apresentam com diferentes alternativas nas seqüências e podem ser encontrados no genoma de indivíduos isoladamente ou grupos de indivíduos, em algumas populações (Figura 2).

```
... GGGCAACTCCAG...   ... GGGCAACTCCAG...   ... GGGCAACTCCAG...  
... GGGCAACTCCAG...   ... GGGCAACTCCAG...   ... GGGCAACTCCAG...
```

Figura 2: Exemplos hipotéticos de SNPs bi, tri e tetra-alélicos, respectivamente. A primeira linha, em negrito, representa a seqüência consenso e as bases sublinhadas, os SNPs.

O que difere um indivíduo dos demais da sua espécie é o seu código genético, ou seja, em sua essência, as seqüências de nucleotídeos que formam as moléculas e seqüências de DNA, RNA e proteínas, que, por sua vez, interagem e formam as células, as quais também, interagem e formam os tecidos, os órgãos, até que, finalmente, formam os indivíduos. Essa é a importância dos SNPs, pois, em síntese, a alteração de um único nucleotídeo, uma única base, em uma dada seqüência, pode alterar a produção de uma certa proteína e, se for o caso, o conjunto dessas alterações pode provocar variações nas características dos indivíduos.

3.1 O *fuzzyMorphic.pl*

O *fuzzyMorphic.pl* é uma plataforma para desenvolvimento de SInS, que faz parte dos trabalhos da tese “Modelos Computacionais para Identificação de Informação Genômica Associada à Resistência ao Carrapato Bovino” (Arbex, 2009c), com o qual é possível implementar SInS para diversos problemas e modelos.

Implementado em Perl, o *fuzzyMorphic.pl* foi desenvolvido como uma ferramenta que permite a modelagem e a implementação de SInS, para os quais fosse possível:

1. para a fuzzificação, representar as funções de pertinência sobre formatos de conjuntos padrão - funções Z, trapezoidal, triangular e S;
2. para a implementação da máquina de inferência, utilizar os modelos de Mamdani ou de Larsen;
3. para a defuzzificação, representar as funções de saída sobre formatos de conjuntos padrão e utilizar o centro dos máximos como método de defuzzificação.

Assim, assumindo essas condições, a partir de um simples arquivo texto com diretivas de descrição dos dados de entrada e dos elementos do modelo do sistema, o *fuzzyMorphic.pl* pode fazer a mineração dos dados e inferir conhecimento, a partir das regras de inferência descritas por diretivas definidas pelos próprios usuários.

No *fuzzyMorphic.pl*, ressaltam-se dois aspectos: inicialmente, como não foi desenvolvido para nenhum problema em específico, permite sua utilização no desenvolvimento de SInS para variados problemas e modelos de inferência nebulosa. O segundo aspecto trata da sua forma de utilização – isto é, seu recurso de descrição e implementação de modelos nebulosos – pois, a facilidade com que permite a descrição de modelos nebulosos, possibilita que um estudo seja realizado sob diferentes condições, com pouco esforço, o que é particularmente necessário e importante em procedimentos de pesquisa, tais como, modelagem e simulação de sistemas físicos, mineração de dados para descoberta de conhecimento ou para tomada de decisão, entre outros.

Sua versão atual, foi desenvolvida e testada sob o sistema operacional GNU/Linux, distribuição Ubuntu, utilizando o GNOME como a interface gráfica para o usuário. Entretanto, por não utilizar nenhum recurso específico, acredita-se que seu funcionamento deva ocorrer em qualquer ambiente Unix-like, contanto que esse possua um ambiente gráfico e o interpretador Perl 5.0 ou posterior. Também faz parte do *fuzzyMorphic.pl* o módulo *fuzzyInference.pm*, que traz a implementação das rotinas e estruturas básicas para a fuzzificação, a inferência e a defuzzificação.

3.2 Modelagem nebulosa para descoberta de SNPs

A descoberta de SNPs por algoritmos computacionais é uma prática bastante difundida e, nessa área, dois programas se destacam pelo amplo uso: Polyphred (Nickerson et al., 1997) e Polybayes (Marth et al., 1999).

O Polyphred, analisa diretamente os sinais expressos no seqüenciamento do material genético e detecta SNPs a partir da variação dos sinais de fluorescência dos cromatogramas, procurando por reduções nas regiões do pico do sinal. Se for encontrada uma redução, onde uma segunda base foi detectada, então esse ponto é identificado como potencial heterozigoto. Após o alinhamento das seqüências (*reads*), as bases dessa seção transversal, que inclui *reads* e consenso, são comparadas. O Polybayes analisa as bases geradas a partir da “leitura” dos cromatogramas - feita por *base-calling* (Ewing et al., 1998), que nomeia e atribui um valor de qualidade para cada base (*Phred quality score* - PQS) - e utiliza um algoritmo de inferência Bayesiana, que procura por seções transversais onde os *reads* alinhados apresentam bases diferentes entre si.

O Polybayes considera o número de *reads* e, ainda, a taxa *a priori* de pontos polimórficos, como sendo $(\frac{1-0.003}{4})$, ou seja, um SNP para cada 333 pares de bases, dividido pelo número de possíveis diferentes bases - A, T, C ou G - em um ponto. Deve ser notado que, esses dois *scripts*, têm seus resultados influenciados pelo PQS, obtido durante a leitura dos cromatogramas.

Os referidos programas trabalham com diferentes metodologias, sobre diferentes atributos, contudo, espera-se que apresentem resultados similares, ao tratarem um mesmo conjunto de dados, mas, não é incomum fornecerem resultados diferentes, o que produz incerteza na tomada de decisão, quando os resultados são discordantes.

Portanto, o modelo apresentado, baseia-se em lógica nebulosa para, a partir dos resultados do Polyphred e do Polybayes, auxiliar na tomada de decisão, no caso em que as informações sejam divergentes e, também, na confirmação de informações coincidentes. Ou seja, utiliza a lógica nebulosa para dar suporte à decisão, avaliando os resultados gerados por dois diferentes métodos e, ainda, incluindo, explicitamente, o PQS das bases do consenso, como um “valorizador” adicional, que reduz os efeitos específicos de cada um dos *scripts*.

A metodologia aqui apresentada não define nenhum limiar de “corte”, no que se refere ao PQS, pois, o modelo de inferência nebulosa, automaticamente, elimina os pontos de baixa qualidade, não classificando-os como SNPs. Os critérios para a definição das variáveis lingüísticas (conjuntos nebulosos), seus qualificadores e das funções de pertinência (Expressões 1 a 9), fundamentaram-se:

1. no *Polyphred score (PPS)*, que estabelece seis classes com intervalos precisos (Nickerson et al., 2008), variando de 1 ($PPS \geq 99$) com uma taxa de verdadeiros positivos de 97%, sendo altamente provável a existência de SNPs; até 6 ($PPS \leq 49$) com uma taxa de verdadeiros positivos de 1%, sendo improvável a existência de SNPs. Logo, a sua função de pertinência foi definida pela variável lingüística *probabilidade*, com os termos (Expressões 1, 2, 3, 4, 5 e 6): *improvável* (P_{IM}), *pouco provável* (P_{PP}), *medianamente provável* (P_{mP}), *provável* (P_{PR}), *muito provável* (P_{MP}) e *altamente provável* (P_{AP});

$$P_{IM}(x) = \begin{cases} 1 & x \leq 49 \\ \frac{59-x}{59-49} & 49 < x < 59 \\ 0 & x \geq 59 \end{cases} \quad (1)$$

$$P_{PP}(x) = \begin{cases} 0 & x \leq 25 \\ \frac{x-25}{50-25} & 25 < x < 50 \\ 1 & 50 \leq x \leq 69 \\ \frac{79-x}{79-69} & 69 < x < 79 \\ 0 & x \geq 79 \end{cases} \quad (2)$$

$$P_{mP}(x) = \begin{cases} 0 & x \leq 60 \\ \frac{x-60}{70-60} & 60 < x < 70 \\ 1 & 70 \leq x \leq 89 \\ \frac{91,5-x}{91,5-89} & 89 < x < 91,5 \\ 0 & x \geq 91,5 \end{cases} \quad (3)$$

$$P_{PR}(x) = \begin{cases} 0 & x \leq 80 \\ \frac{x-80}{90-80} & 80 < x < 90 \\ 1 & 90 \leq x \leq 94 \\ \frac{96-x}{96-94} & 94 < x < 96 \\ 0 & x \geq 96 \end{cases} \quad (4)$$

$$P_{MP}(x) = \begin{cases} 0 & x \leq 92,5 \\ \frac{x-92,5}{95-92,5} & 92,5 < x < 95 \\ 1 & 95 \leq x \leq 98 \\ \frac{99-x}{99-98} & 98 < x < 99 \\ 0 & x \geq 99 \end{cases} \quad (5)$$

$$P_{AP}(x) = \begin{cases} 0 & x \leq 96,5 \\ \frac{x-96,5}{99-96,5} & 96,5 < x < 99 \\ 1 & x \geq 99 \end{cases} \quad (6)$$

2. na qualidade das bases do consenso – o *Phrap quality score (PQS)* – que varia entre 4 e 90 e é separada, pelo limiar $PQS = 20$, em duas classes de valores e sua função de pertinência (Expressões 7, 8 e 9), foi definida como a variável lingüística *qualidade*, nos termos: ruim (Q_R), boa (Q_B) e ótima (Q_O).

$$Q_R(x) = \begin{cases} 1 & x \leq 20 \\ \frac{30-x}{30-20} & 20 < x < 30 \\ 0 & x \geq 30 \end{cases} \quad (7)$$

$$Q_B(x) = \begin{cases} 0 & x \leq 15 \\ \frac{x-15}{30-15} & 15 < x < 30 \\ 1 & 30 \leq x \leq 40 \\ \frac{70-x}{70-40} & 40 < x < 70 \\ 0 & x \geq 70 \end{cases} \quad (8)$$

$$Q_O(x) = \begin{cases} 0 & x \leq 40 \\ \frac{x-40}{50-40} & 40 < x < 50 \\ 1 & x \geq 50 \end{cases} \quad (9)$$

As classificações determinadas pelo *PPS* (Tabela 2) e o pelo *PQS* são propostas a partir de abordagens clássicas da lógica e da matemática, com limites precisos e, assim, sob raciocínio incerto ou impreciso podem levar a resultados matematicamente corretos, porém distorcidos, como a situação proposta em 2.1.

Tabela 2: Classes definidas pelo PPS.

Classc	PPS	Taxa de verdadeiros positivos
1	99	97%
2	95 - 98	75%
3	90 - 94	62%
4	70 - 89	35%
5	50 - 69	11%
6	0 - 49	1%

Assim, nesse modelo de inferência proposto, os valores discretos de entrada - os PPSs, encontrado pelo *Polyphred* e pelo seu equivalente no *Polybayes*, e o *PQS* - têm seus graus de pertinência estabelecidos pelas Expressões 1 a 9, que “disparam” regras nebulosas, cujo resultado é discretizado pelo método do “centro dos máximos” (*center of máxima* - CoM) – ou “defuzzificação pelas alturas” (*height defuzzification* - HD) – visto que esse considera a ocorrência de múltiplos disparos de regras sobre uma mesma saída, “valorizando” essa saída. Como resultado, determina-se um novo valor, mais apurado, indicativo da existência de polimorfismo, para cada SNP anteriormente identificado, onde foram considerados os valores iniciais dos PPSs e da *PQS* no ponto.

3.3 Análise do modelo e do sistema de inferência nebulosa

O modelo nebuloso investiga o conjunto de dados originado a partir da junção dos conjuntos gerados pelo *Polyphred* e pelo *Polybayes*; avalia as probabilidades, estabelecidas por suas diferentes propostas, de cada elemento do conjunto; e, então, determina para cada um dos elementos um novo valor, para servir como uma referência na tentativa de se particionar resultados em grupos.

Ou seja, o SIN inclui um novo atributo, o *PQS*, ampliando as possibilidades de investigação e utiliza-se deste como um “valorizador” para a tomada da decisão. Assim, aos resultados prévios de o ponto vir a ser um SNP, acrescenta-se a qualidade do ponto, utilizando-os como as variáveis permitem a determinação de uma das três possibilidade excludentes: a confirmação do SNP, a eliminação dessa possibilidade ou, uma situação intermediária, sem elementos conclusivos para a confirmação ou o descarte dessa possibilidade.

Resultados efetivos desse modelo são obtidos mediante a análise de conjuntos de dados, quando os valores inferidos, podem ser agrupados, determinando os conjuntos de pontos que melhor se ajustam às possibilidades investigadas. Estabelecer grupos é uma tarefa complexa, pois procura-

se dizer como são e em quantas classes os dados se distribuem, sem o conhecimento *a priori* dos mesmos e, caso os valores se distribuam equitativamente no espaço, as classes podem não existir, uma vez que são definidas com base na semelhança entre os elementos, cabendo a verificação das possíveis classes para avaliar a existência de algum significado útil (de Carvalho, 2005).

Entretanto, qualquer classificação que se queira fazer pode vir a ser influenciada pela forma ou comportamento dos dados ou, ainda, para classes definidas com limites precisos, promover decisões duvidosas para valores limítrofes. Tais questões, entre outras, sugerem a adoção de métodos de particionamento não-hierárquicos e não-supervisionados, pois não partem de nenhuma premissa externa para estabelecer as classes, mas, de forma oposta, suas premissas são estabelecidas por características específicas, internas e inerentes ao conjunto avaliado, eliminando ou reduzindo a ação de agentes externos ao modelo, como a definição *a priori* de limites precisos para as classes.

4 Conclusões

Critérios fixos e precisos de classificação, em geral, não são adequados, quando a análise dos dados apresenta resultados que situam-se próximos à divisão das classes, o que pode ser tratado por SINS e que também são propostos para a problemas que apresentam incerteza ou imprecisão.

Nesse sentido, o *fuzzyMorphic.pl* é uma ferramenta de implementação de SINS de simples utilização, bastando descrever as variáveis de entrada, as funções de pertinência, as regras de inferência, a máquina de inferência e a função de saída.

Ao adicionar um novo atributo aos resultados prévios, o sistema nebuloso é capaz de decidir, de forma única, entre suas três possibilidades e, então, ao agrupá-las a partir de um algoritmo não-supervisionado e com estabelecimento dinâmico do número de grupos, espera-se que o resultado desse agrupamento estabeleça em três grupos, não necessitando de limites fixos e precisos para a identificação de possíveis SNPs.

Referências

- Arbex, W. (2009a). *fuzzyMorphic.pl*, 1 CD. Perl. Ambiente UNIX-like com GUI e interpretador Perl 5.0 ou posterior.
- Arbex, W. (2009b). *Introdução ao fuzzy-Morphic.pl 1.0*, 1 edn, Embrapa Gado de Leite, Juiz de Fora. Manual do *fuzzy-Morphic.pl* Versão 1.0-20090111.
- Arbex, W. (2009c). *Modelos computacionais para identificação de informação genômica associada à resistência ao carrapato bovino*, Doutorado em Engenharia de Sistemas e Computação, Universidade Federal do Rio de Janeiro, Rio de Janeiro. Trabalho vencedor do prêmio SBI Agro 2009 - Categoria Tese de Doutorado no VII Congresso da Sociedade Brasileira de Agroinformática.
- Cox, E. (1994). *The fuzzy systems handbook: a practitioner's guide to building, using, and maintaining fuzzy systems*, Academic Press Professional, Inc., San Diego, CA, USA.
- de Almeida, P. E. M. and Evsukoff, A. G. (2005). *Sistemas fuzzy*, Manole, Barueri, pp. 169–202.
- de Carvalho, L. A. V. (2005). *Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração*, Ciência Moderna, Rio de Janeiro.
- Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998). Base-calling of automated sequencer traces using Phred (I): accuracy assessment, *Genome Research* **8**(3): 175–185.
- Goldschmidt, R. and Passos, E. (2005). *Data mining: um guia prático*, Elsevier, Rio de Janeiro.
- Klir, G. J. and Yuan, B. (1995). *Fuzzy sets and fuzzy logic: theory and applications*, Prentice Hall, Upper Saddle River.
- Marth, G. T., Korf, I., Yandell, M. D., Yeh, R. T., Gu, Z., Zakeri, H., Stitzel, N. O., Hillier, L., Kwok, P.-Y. and Gish, W. R. (1999). A general approach to single-nucleotide polymorphism discovery, *Nature Genetics* **23**: 452–456.
- McNeill, F. M. and Thro, E. (1994). *Fuzzy logic: a practical approach*, Academic Press Professional, Inc., San Diego, CA, USA.
- Nickerson, D. A., Taylor, S. L., Kolker, N., Sloan, J., Bhangale, T., Stephens, M. and Robertson, I. (2008). *Polyphred users manual*, University of Washington, Seattle. Version 6.15 Beta.
- Nickerson, D. A., Tobe, V. O. and Taylor, S. L. (1997). PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing, *Nucl. Acids Res.* **25**(14): 2745–2751.
- Tanscheit, R. (2007). *Sistemas fuzzy*, Thomson Learning, São Paulo, pp. 229–264.
- Zadeh, L. A. (1965). Fuzzy sets, *Information and Control* **8**(3): 338–353.
- Zadeh, L. A. (1973). Outline of a new approach to the analysis of complex systems and decision processes, *IEEE Trans. on Systems, Man, and Cybernetics* **SMC-3**: 28–44.