Inferência *Fuzzy* para Suporte à Decisão na Descoberta de SNPs

Wagner Arbex, Marta Martins, Marcos Vinícius Silva
Centro Nacional de Pesquisa de Gado de Leite
Empresa Brasileira de Pesquisa Agropecuária
Juiz de Fora, MG, Brasil
{arbex, mmartins, marcos}@cnpgl.embrapa.br

Luís Alfredo Carvalho
Programa de Engenharia de Sistemas de Computação
Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ, Brasil
LuisAlfredo@ufrj.br

Resumo—**Diferenças pontuais entre pares de bases de diferentes sequências alinhadas são o tipo mais comum de variabilidade genética e, sendo conhecidas como polimorfismos de base única (*single nucleotide polymorphisms* - SNPs), são importantes no estudo da variabilidade das espécies, pois podem provocar alterações funcionais ou fenotípicas, que, por sua vez, podem implicar em consequências evolutivas ou bioquímicas nos indivíduos das espécies. A identificação de SNPs é feita por diversos modelos matemáticos e computacionais e esse trabalho apresenta um modelo que utiliza a lógica *fuzzy* como base para o desenvolvimento 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.**

Palavras-chave–inferência fuzzy; suporte à decisão; modelagem matemática; modelagem computacional; single nucleotide polymorphisms.

I. INTRODUÇÃO

A descoberta de SNPs por algoritmos computacionais é uma prática bastante difundida e, assim, existem diversas ferramentas que implementam diferentes metodologias, sobre diferentes atributos. Contudo, espera-se que, se comparadas, apresentem resultados similares, ao tratarem um mesmo conjunto de dados de entrada que, nesse caso, em geral, esses dados são sequências genômicas.

Entretanto, não é incomum que diferentes ferramentas forneçam resultados diferentes, o que produz incerteza na tomada de decisão, quando os resultados são discordantes.

O presente texto apresenta um modelo que se baseia em lógica *fuzzy* para, a partir de resultados prévios, auxiliar na tomada de decisão, no caso em que as informações sejam divergentes e, também, na confirmação de informações coincidentes.

Ou seja, o modelo desenvolvido, utiliza a lógica *fuzzy* para dar suporte à decisão, a partir de resultados gerados por dois diferentes métodos e, ainda, incluindo, explicitamente, um atributo adicional, como "valorizador" adicional, que reduz os efeitos específicos de cada um dos métodos que geraram os resultados prévios.

II. SINGLE NUCLEOTIDE POLYMORPHISMS

Uma das variações e particularidades do genoma da maioria das espécies são os chamados polimorfismos de base única (*single nucleotide polymorphisms* – SNPs), modificações de um único nucleotídeo, em uma dada sequência, quando comparada a outra. Ou seja, SNPs são pares de bases em uma única posição do DNA genômico que se apresentam com diferentes alternativas nas sequências (Fig. 1) – isto é, alelos – e podem ser encontrados no genoma de indivíduos normais em algumas populações ou grupos de indivíduos.

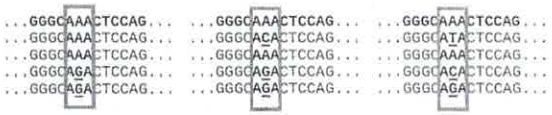

Figura 1. Exemplos hipotéticos de SNPs bi, tri e tetra-alélicos, respectivamente. A primeira linha, em negrito, representa a sequência consenso e as bases sublinhadas, os SNPs. Na prática, a ocorrência de SNPs bi-alélicos não é somente mais comum, mas, quase absoluta em relação as demais formas [1].

A maior parte do genoma dos indivíduos de uma mesma espécie é idêntica, porém, existe a variabilidade genética, que consiste na alteração nas sequências de bases ao longo do DNA. Tais alterações ocorrem por substituição, ausência ou duplicação de bases e os SNPs são o tipo mais comum de variabilidade genética [2].

O que difere um indivíduo dos demais da sua espécie é o código genético, isto é, as sequências de nucleotídeos que formam as moléculas e sequências de DNA, RNA e proteínas, que, por sua vez, interagem e formam as células, as quais, por sua vez, formam os tecidos, os órgãos, até que, finalmente, formam os indivíduos. Ou seja, as diferenças se iniciam na ordem em que os nucleotídeos se apresentam e essa é a importância dos SNPs, já que a alteração de um único nucleotídeo em uma dada sequência pode alterar a produção de uma certa proteína.

Assim, SNPs são importantes no estudo da variabilidade das espécies, uma vez que podem provocar alterações funcionais ou fenotípicas, que podem implicar em consequências evolutivas ou bioquímicas nos indivíduos em que esses polimorfismos se manifestam.

III. INFERÊNCIA *FUZZY* COMO SUPORTE À DECISÃO

A subjetividade no raciocínio em geral, utilizada no cotidiano, sendo transmitida e perfeitamente compreendida entre interlocutores, é expressa em "termos e variáveis linguísticas" [3] e não é expressa sob a lógica clássica ou qualquer abordagem matemática tradicional. O uso de, por exemplo, adjetivos comuns que representam imprecisão ou incerteza, tais como, *alto*, *baixo* ou, 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 determine a altura, a partir da qual, uma pessoa pode ser considera alta.

Os termos e variáveis linguísticas aumentam a complexidade dos sistemas computacionais frente à capacidade de trabalharem com números, valores exatos, discretos e, por vezes, excludentes, o que sugere a ideia de que, trabalhar com valores incertos, possibilita a modelagem de sistemas complexos, mesmo que se reduza a precisão do resultado, mas não retira a credibilidade.

Se as incertezas, quando consideradas isoladamente, são indesejáveis, quando associadas a outras características, em geral, permitem a redução da complexidade do sistema e aumentam a credibilidade dos resultados obtidos [4].

As abordagens clássicas são falhas para valores limítrofes e, portanto, resultados matemática e logicamente precisos, porém, questionáveis, podem ser encontrados. Por exemplo, o estudo de caso a ser apresentado na Seção IV, será utilizado o *polyphred score* (PPS) que estabelece seis classes com intervalos precisos (Tab. I) [5] e, supondo que fossem determinados os *scores* 70 e 89 para dois pontos, então, para ambos, seria considerada a taxa de 35% de verdadeiros positivos na decisão desses pontos virem a ser SNPs (Classe 4).

TABELA I. CLASSES DEFINIDAS PELO PPS

Classe	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%

Essa decisão, lógica e matematicamente precisa, pode ser questionada devido à subjetividade que a envolve, visto que, 70 e 89, se encontram nos extremos da classe a qual pertencem e, portanto, muito próximos de diferentes interpretações.

Todavia, as abordagens clássicas da lógica e da matemática não possuem as ferramentas necessárias para tratar valores limítrofes, imprecisão ou incerteza. Um valor limítrofe acarretará dúvidas na "decisão" de o ponto ser, ou não, considerado polimórfico, o que sugere um modelo de inferência *fuzzy* para o tratamento dessa incerteza.

O problema de valores limítrofes, em geral, não é tão simples quanto parece, do contrário, as abordagens clássicas poderiam facilmente resolvê-lo, mas, ao aproximar-se do raciocínio subjetivo para a interpretação e a extração de uma resposta, uma decisão, torna-se complexo e a aparente simplicidade é conferida pela modelagem por lógica *fuzzy* e seu embasamento na teoria dos conjuntos *fuzzy*.

A subjetividade intrínseca ao raciocínio trata situações complexas, mediante imprecisão, incerteza ou aproximação e, então, são utilizados "operadores humanos", também de natureza imprecisa, que são expressos por termos ou variáveis linguísticas, o que, em geral, não permite uma solução em termos exatos, mas, pode propor uma classificação, agrupamento ou agregação qualitativa em categorias ou possíveis conjuntos de soluções [6].

IV. DESCRIÇÃO DO MODELO DE INFERÊNCIA *FUZZY* PARA SUPORTE À DECISÃO

No modelo proposto, consideram-se como valores de entrada, as probabilidades, previamente determinadas, de o ponto vir a ser um SNP e o valor de qualidade do ponto na sequência consenso.

Os *Casos 1* e *2* serão utilizados ao longo do texto para demonstrar o modelo, assumindo, para o *Caso 1*, 99% e 96%, quanto às probabilidades e 43 de qualidade e, para o *Caso 2*, os valores são, respectivamente, 94%, zero e 50.

Esses casos são parte de um projeto de pesquisa [7], que desenvolveu, implementou e executou o modelo de decisão em discussão, sobre 4072 sequências expressas identificadas (*expressed sequence tags* – EST) relacionadas à expressão de resistência de bovinos à ação do "carrapato do boi", na busca de informações genômicas, em específico, SNPs que estivessem associados à resistência dos bovinos a esses ácaros.

Os valores das probabilidades, que, a princípio, deveriam ser classificados diretamente na Tab. I, em uma tentativa de se identificar um SNP, foram obtidos com o uso dos programas Polyphred [8] e Polybayes [9].

A descoberta de SNPs por algoritmos computacionais é uma prática bastante difundida e os programas Polyphred e Polybayes se destacam pelo seu amplo uso nessa área.

Esses programas possuem diferentes métodos para a obtenção de seus resultados e podem apresentar valores muito conflitantes, como o exemplificado no *Caso 2*. Assim, uma comparação simples desses resultados com a Tab I, pode levantar ainda mais dúvidas, quando o que se buscava era uma resposta "exata" e, além disso, deve ser notado que, esses dois programas, têm seus resultados influenciados pelo escore de qualidade da base *phred quality score* (PQS), obtido durante a leitura dos cromatogramas.

Em geral, a estrutura de um modelo de inferência *fuzzy* é organizada em três etapas:

- a fuzzificação, para converter os valores precisos em valores *fuzzy*;
- a inferência, que executa a "máquina" de inferência co as regras de inferência;
- a defuzzificação, para converter os valores de saída, de valores fuzzy em valores precisos.

Essa estrutura, é apresentada por inúmeros autores, com pouca ou nenhuma alteração e suas etapas são apresentadas no escopo do caso discutido.

A. Fuzzificação

Avalia-se um valor de entrada por sua "função de pertinência", o que determina um "grau de pertinência" (GP) do valor para a sua função e as funções de pertinência adotadas foram baseadas:

1. no PPS (Tab. I), com a função de pertinência definida pela variável linguística *probabilidade*, com os termos (Eqs. 1): *improvável* (P_{IM}), *pouco provável* (P_{PP}), *medianamente provável* (P_{mP}), *provável* (P_{PR}), *muito provável* (P_{MP}) e *altamente provável* (P_{AP});

$$P_{IM}(X) = \begin{cases} 1 & x \le 49 \\ \dfrac{59 - x}{59 - 49} & 49 < x < 59 \\ 0 & x \ge 59 \end{cases}$$

$$P_{PP}(X) = \begin{cases} 0 & x \le 25 \\ \dfrac{x - 25}{50 - 25} & 25 < x < 50 \\ 1 & 50 \le x \le 69 \\ \dfrac{79 - x}{79 - 69} & 69 < x < 79 \\ 0 & x \ge 79 \end{cases}$$

$$P_{mP}(X) = \begin{cases} 0 & x \le 60 \\ \dfrac{x - 60}{70 - 60} & 60 < x < 70 \\ 1 & 70 \le x \le 89 \\ \dfrac{91{,}5 - x}{91{,}5 - 89} & 89 < x < 91{,}5 \\ 0 & x \ge 91{,}5 \end{cases} \quad (1)$$

$$P_{PR}(X) = \begin{cases} 0 & x \le 80 \\ \dfrac{x - 80}{90 - 80} & 80 < x < 90 \\ 1 & 90 \le x \le 94 \\ \dfrac{96 - x}{96 - 94} & 94 < x < 96 \\ 0 & x \ge 96 \end{cases}$$

$$P_{MP}(X) = \begin{cases} 0 & x \le 92{,}5 \\ \dfrac{x - 92{,}5}{95 - 92{,}5} & 92{,}5 < x < 95 \\ 1 & 95 \le x \le 98 \\ \dfrac{99 - x}{99 - 98} & 98 < x < 99 \\ 0 & x \ge 99 \end{cases}$$

$$P_{AP}(X) = \begin{cases} 0 & x \le 96{,}5 \\ \dfrac{x - 96{,}5}{99 - 96{,}5} & 96{,}5 < x < 99 \\ 1 & x \ge 99 \end{cases}$$

2. na qualidade das bases do consenso – o PQS -- que varia entre 4 e 90 e sua função de pertinência define a variável linguística *qualidade*, nos termos (Eqs. 3): *ruim* (Q_R), *boa* (Q_B) e *ótima* (Q_O).

$$Q_R(X) = \begin{cases} 1 & x \le 20 \\ \dfrac{30 - x}{30 - 20} & 20 < x < 30 \\ 0 & x \ge 30 \end{cases}$$

$$Q_B(X) = \begin{cases} 0 & x \le 15 \\ \dfrac{x - 15}{30 - 15} & 15 < x < 30 \\ 1 & 30 \le x \le 40 \\ \dfrac{70 - x}{70 - 40} & 40 < x < 70 \\ 0 & x \ge 70 \end{cases} \quad (2)$$

$$Q_O(X) = \begin{cases} 0 & x \le 40 \\ \dfrac{x - 40}{50 - 40} & 40 < x < 50 \\ 1 & x \ge 50 \end{cases}$$

Os resultados da fuzzificação para o *Caso 1*, $PPS_{PP} = 99$, $PPS_{PB} = 96$ e $PQS = 43$, em suas respectivas funções de pertinência, podem ser vistos nas Tabs. II e III e, para o *Caso 2*, o resultado da fuzzificação para $PPS_{PP} = 94$, $PPS_{PB} = 0$ e $PQS = 50$, pode ser visto nas Tabs. IV e V.

TABELA II. GPS PARA A *PROBABILIDADE* (*CASO 1*).

	PPS_{PP}	PPS_{PB}
Improvável	0	0
Pouco provável	0	0
Medianamente provável	0	0
Provável	0	0
Muito provável	0	1
Altamente provável	1	0

TABELA III. GPS PARA A *QUALIDADE* (*CASO 1*).

	PQS
Ruim	0
Boa	0,9
Ótima	0,3

TABELA IV. GPS PARA A PROBABILIDADE (CASO 2).

	PPS_{PP}	PPS_{PB}
Improvável	0	1
Pouco provável	0	0
Medianamente provável	0	0
Provável	1	0
Muito provável	0,6	0
Altamente provável	0	0

TABELA V. GPS PARA A *QUALIDADE* (*CASO 2*).

	PQS
Ruim	0
Boa	0,67
Ótima	1

A *probabilidade*, para o *Caso 1*, é expressa pelos termos *muito provável* e *altamente provável*, e a *qualidade*, pelos termos *bom* e *ótimo* e, essas mesmas variáveis do *Caso 2*, pelos termos *improvável*, *provável*, *muito provável*, *bom* e *ótimo*.

B. Inferência

A inferência executa operações sobre os conjuntos *fuzzy*, com a combinação dos antecedentes das regras, a implicação e a aplicação do *modus ponens* generalizado, sendo, esse procedimento, feito em dois passos: a "agregação", que corresponde ao operador lógico *E* que executa a intersecção entre conjuntos e, portanto, determina o mínimo entre os valores disparados pelas regas, seguido da "composição".

Os modelos ("máquinas") de inferência adequados para o caso em questão são os modelos de Mamdani [10] ou de Larsen [11], visto que são sensíveis ao disparo de múltiplas regras sobre o conjunto de saída, quando, então, inicia-se o procedimento de defuzzificação, que começa com o segundo passo da inferência, a "composição", que é equivalente ao operador lógico *OU* e executa a união entre conjuntos, na qual o maior valor entre os mínimos resultantes da agregação é considerado para a defuzzificação.

Foram estabelecidas trinta e seis regras de inferência (Fig. 2), sendo que, em metade dessas, seus antecedentes são avaliados pelas variáveis *probabilidade* (PPS_{PP}) e *qualidade* e, a outra metade, é avaliada pelas variáveis *probabilidade* (PPS_{PB}) e *qualidade*.

	Q_R		P_{IM}		SNP_D	(R_1)
	Q_R		P_{PP}		SNP_D	(R_2)
	Q_R		P_{mP}		SNP_D	(R_3)
	Q_R		P_P		SNP_D	(R_4)
	Q_R		P_{MP}		SNP_D	(R_5)
	Q_R		P_{AP}		SNP_D	(R_6)
	Q_B		P_{IM}		SNP_D	(R_7)
	Q_B		P_{PP}		SNP_D	(R_8)
SE	Q_B	E	P_{mP}	ENTÃO	SNP_{NC}	(R_9)
	Q_B		P_P		SNP_{NC}	(R_{10})
	Q_B		P_{MP}		SNP_C	(R_{11})
	Q_B		P_{AP}		SNP_C	(R_{12})
	Q_O		P_{IM}		SNP_D	(R_{13})
	Q_O		P_{PP}		SNP_D	(R_{14})
	Q_O		P_{mP}		SNP_{NC}	(R_{15})
	Q_O		P_P		SNP_{NC}	(R_{16})
	Q_O		P_{MP}		SNP_C	(R_{17})
	Q_O		P_{AP}		SNP_C	(R_{18})

Figura 2. Regras de inferência utilizadas no modelo

Essas regras relacionam termos de entrada com a função de saída, expressa pelos termos *SNP descartado* (SNP_D), *SNP não confirmado* (SNP_{NC}) e *SNP confirmado* (SNP_C) e os possíveis resultados da aplicação das regras de inferência estão representados na Tab. VI.

TABELA VI. DECISÕES A PARTIR DAS REGRAS DE INFERÊNCIA

	P_{IM}	P_{PP}	P_{mP}	P_{PR}	P_{MP}	P_{AP}
Q_R	SNP_D	SNP_D	SNP_D	SNP_D	SNP_D	SNP_D
Q_B	SNP_D	SNP_D	SNP_{NC}	SNP_{NC}	SNP_C	SNP_C
Q_O	SNP_D	SNP_D	SNP_{NC}	SNP_{NC}	SNP_C	SNP_C

No *Caso 1*, as funções de pertinência (Eqs. 1 e 2), resultam em $P_{MP} = 1$, para PPS_{PB}, $P_{AP} = 1$, para PPS_{PP}, $Q_B = 0{,}9$ e $Q_O = 0{,}3$ (Tabs. II e III), então, a agregação é feita entre Q_B e Q_O, o que resulta no termo *ótima* para a variável *qualidade*. Os demais valores obtidos são iguais e, assim, não aplica-se a agregação, o que resulta em *muito provável* (PPS_{PB}) e *altamente provável* (PPS_{PP}), para *probabilidade*, que disparam as regras R_{17} e R_{18}.

Para o *Caso 2*, após a agregação, toma-se $P_{IM} = 1$ (PPS_{PB}), $P_{MP} = 0{,}6$ (PPS_{PP}) e $Q_B = 0{,}67$ (Tabs. II e III) que são levados à máquina de inferência, que dispara R_7 e R_{11}.

O modelo de inferência mapeia os antecedentes, resultantes da agregação, no termo consequente, que, para os modelos de Mamdani ou Larsen, representa uma função de saída em termos linguísticos, exatamente como uma função de pertinência.

A função de saída que foi estabelecida, reduz as seis classes definidas para o PPS aos termos *SNP descartado*, *SNP não confirmado* e *SNP confirmado*, que, então, compõem a variável linguística *SNP* (Eqs. 3):

$$SNP_D(X) = \begin{cases} 1 & x \le 20 \\ \dfrac{30 - x}{30 - 20} & 20 < x < 30 \\ 0 & x \ge 30 \end{cases}$$

$$SNP_{NC}(X) = \begin{cases} 0 & x \le 15 \\ \dfrac{x - 15}{30 - 15} & 15 < x < 30 \\ 1 & 30 \le x \le 40 \\ \dfrac{70 - x}{70 - 40} & 40 < x < 70 \\ 0 & x \ge 70 \end{cases} \qquad (3)$$

$$SNP_C(X) = \begin{cases} 0 & x \le 40 \\ \dfrac{x - 40}{50 - 40} & 40 < x < 50 \\ 1 & x \ge 50 \end{cases}$$

As regras R_{17} e R_{18}, disparadas no *Caso 1*, são processadas como:

- R_{17} tem como antecedentes o valor *muito provável*, com GP = 1, e o valor *ótima*, com GP = 0,3; assim, a aplicação da regra mapeia o consequente *SNP confirmado*, com GP = 1 e GP = 0,3, isto é $SNP_C = 1$ e $SNP_C = 0{,}3$;
- R_{18} tem como antecedentes o valor *altamente provável*, com GP = 1, e o valor *ótima*, com GP = 0,3; então, da mesma forma, mapeia o consequente *SNP confirmado*, com GP = 1 e GP = 0,3, isto é $SNP_C = 1$ e $SNP_C = 0{,}3$.

Com a aplicação das duas regras, cujos resultados foram coincidentes, apenas o termo *SNP confirmado* foi mapeado e o procedimento de composição deve ser tomado somente sobre esse termo. A composição busca o máximo entre os GPs de cada termo, no caso, somente sobre o termo *SNP confirmado*, fazendo $SNP_C = 1$.

Para o *Caso 2*, são disparadas as regras R_7 e R_{11}, que avaliam os valores antecedentes $P_{IM} = 1$ e $Q_O = 0{,}67$, para R_7, e $P_{MP} = 0{,}6$ e $Q_O = 0{,}67$, para R_{11}. A regra R_7 mapeia na função de saída o valor *SNP descartado*, com GP = 1 e GP = 0,67, enquanto a regra R_{11} mapeia na função de saída o valor *SNP não confirmado*, com GP = 0,67 e GP = 0,6. O termo *SNP*

confirmado não foi mapeado, logo o procedimento de composição aplicado aos demais termos resulta em *SNP descartado*, com GP = 1 (SNP_D = 1), e *SNP não confirmado*, com GP = 0,67 (SNP_{NC} = 0,67).

C. *Defuzzificação*

A defuzzificação executa a composição, que determina os valores que representam cada um dos conjuntos mapeados na função de saída, e, a partir desses, calcula um valor preciso (VP), obtido com a aplicação do método de defuzzificação.

Para o modelo proposto, o método de defuzzificação deve considerar múltiplos disparos, pois o valor da qualidade da base no consenso é utilizada como um "valorizador" dos valores de probabilidade confrontados (PPS_{PP} e PPS_{PB}). Assim, havendo disparos múltiplos, esses devem ser avaliados, pois, servem à ideia de valorizar os conjuntos *fuzzy* estabelecidos na função de saída. Para esse fim, deve ser utilizado o método centro dos máximos (*center of maxima* - COM) e, a partir dos modelos de inferência, aplica-se o método de defuzzificação adequado ao problema.

O modelo de inferência em discussão permite a inferência pelas "máquinas" de Mamdani e Larsen, assim, ambos podem ser aplicados e, juntamente com os valores tomados da composição, definem os valores para a defuzzificação.

O COM (Eq. 4), trata-se de uma média ponderada, onde o numerador é o somatório dos valores da composição (h_i), isto é, a altura dos conjuntos de saída, multiplicados pelos valores no universo de discurso (u_i), encontrados pelo modelo de inferência, do seu respectivo conjunto de saída, e o denominador é o somatório das alturas (h_i).

$$VP = \frac{\sum h_i \cdot u_i}{\sum h_i} \qquad (4)$$

Para o *Caso 1*, VP_1 = 75 é igual para os modelos de Mamdani e Larsen, mas, para o *Caso 2*, como consequência desses modelos, a defuzzificação apresenta diferentes resultados VP_2 = 21,02 e VP_2 = 21,02.

V. Discussão Sobre o Modelo de Inferência *Fuzzy*

O *Caso 1* inicia com resultados prévios similares, 99% e 96% de probabilidades do ponto vir a ser um SNP, entretanto, o *Caso 2*, parte de resultados divergentes, 94% e zero.

A modelagem do problema incluiu um novo atributo, a qualidade da base no consenso, 43 e 50, para os *Casos 1* e *2*, respectivamente, ampliando as possibilidades de investigação e utilizando-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 do modelo que permite a determinação de uma das três possibilidades excludentes: a confirmação do SNP (*SNP confirmado*), a eliminação dessa possibilidade (*SNP descartado*) ou, uma situação intermediária, sem elementos conclusivos para a confirmação ou o descarte dessa possibilidade (*SNP não confirmado*).

A análise desses casos fornece elementos suficientes para apresentar o modelo, contudo, resultados efetivos são obtidos mediante a análise de conjuntos de dados, quando os valores inferidos a partir do modelo, podem, então, 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, não caracterizando qualquer categoria, 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 [12].

VI. Considerações Finais

O modelo de aprendizado de máquina apresentado, sob o aspecto funcional, 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 atributo que deve servir como uma referência na tentativa de particionar o conjunto de dados em grupos de elementos que podem ser tratados como *SNP confirmado*, *SNP descartado* e, ainda, *SNP não confirmado*.

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, quando o valor estiver muito próximo dos limites das classes.

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 que podem particionar um conjunto, 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.

As premissas dos métodos de particionamento decorrentes de algoritmos não-hirárquicos baseiam-se no próprio conjunto de valores avaliados, buscando a máxima coesão interna dos objetos de um grupo e o máximo isolamento entre os grupos [12]. Em outras palavras, a partir da análise do próprio conjunto, busca-se identificar os elementos que, em relação ao atributo avaliado, possuem a "menor distância" entre os elementos do grupo e, uma vez estabelecidos grupos cujos elementos possuem essa característica, tais grupos devem apresentar a "maior distância" entre si.

Assim, como essas premissas são decorrentes dos próprios valores analisados, reduzem o efeito de comportamento dos dados. Isto é, supondo que o atributo avaliado apresente uma determinada tendência, todos os elementos possuem o mesmo

comportamento e um particionamento não direcionado e tomado a partir dos próprios elementos pode reduzir ou eliminar essa tendência.

Como visto, 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 modelos de inferência *fuzzy* que também são convenientes e possuem capacidade para a problemas que apresentam incerteza ou imprecisão para a tomada de decisão.

Portanto, com a adição de um novo atributo aos resultados prévios, o modelo *fuzzy* 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

[1] T. Brown, Genomes, New York, John Wiley & Sons, 2 ed., 2002.

[2] The International HapMap Consortium, "The international hapmap pro ject," Nature, vol. 426, pp. 789–796, Dec. 2003.

[3] L. A. Zadeh, "Outline of a new approach to the analysis of complex systems and decision processes," IEEE Trans. on Systems, Man, and Cybernetics, vol. SMC-3, pp. 28–44, 1973.

[4] G. J. Klir and B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications, Upper Saddle River, Prentice Hall, 1995.

[5] D. A. Nickerson, S. L. Taylor, N. Kolker, J. Sloan, T. Bhangale, M. Stephens, and I. Robertson, Polyphred users manual. University of Washington, Seattle, May 2008. Version 6.15 Beta.

[6] P. E. M. de Almeida and A. G. Evsukoff, Sistemas Fuzzy, Barueri, Manole, 2005, pp. 169–202.

[7] W. Arbex, Modelos computacionais para identificação de informação genômica associada à resistência ao carrapato bovino. Tese de Doutorado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Mar. 2009. Programa de Engenharia de Sistemas e Computação.

[8] D. A. Nickerson, V. O. Tobe, and S. L. Taylor, "PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing," Nucl. Acids Res., vol. 25, no. 14, pp. 2745–2751, 1997.

[9] G. T. Marth, I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O. Stitziel, L. Hillier, P.-Y. Kwok, and W. R. Gish, "A general approach to single-nucleotide polymorphism discovery," Nature Genetics, vol. 23, pp. 452–456, Dec. 1999.

[10] E. H. Mamdani, "Application of fuzzy algorithms for control of simple dynamic plant," Proceedings of the Institution of Electrical Engineers, vol. 121, no. 12, pp. 1585–1588, 1974.

[11] P. M. Larsen, "Industrial applications of fuzzy logic control," International Journal of Man-Machine Studies, vol. 12, no. 1, pp. 3 – 10, 1980.

[12] L. A. V. de Carvalho, Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração. Rio de Janeiro: Ciência Moderna, 2005.

Sistemas y Tecnologías de Información

Actas de la 7ª Conferencia Ibérica de Sistemas y Tecnologías de Información
Madrid, España
20 a 23 de Junio de 2012

Vol. I – Artículos

Editores
Álvaro Rocha
Jose A. Calvo-Manzano
Luís Paulo Reis
Manuel Pérez Cota

Artículos de la Conferencia

Artículos de los Workshops

Second Iberian Workshop on Serious Games and Meaningful Play
Third Ibero-American Workshop on Data Quality
Fourth Workshop on Intelligent Systems and Applications
Second Workshop in Information Systems for Interactive Spaces
Second European Workshop on Computing and ICT Professionalism
First Workshop on Information Society and Regional-Urban Development
First Workshop on Computational Biomedical Image Processing and Analysis

aisti
Associação Ibérica de Sistemas e Tecnologias de Informação

POLITÉCNICA

Sistemas y Tecnologías de Información

Actas de la 7ª Conferencia Ibérica de Sistemas y Tecnologías de Información
Madrid, España
20 al 23 de Junio de 2012
AISTI | UPM

Vol. I – Artículos
Tomo 1

Editores
Álvaro Rocha
Jose A. Calvo-Manzano
Luís Paulo Reis
Manuel Pérez Cota

ISBN
978-989-96247-6-4

CRÉDITOS

TÍTULO
Sistemas y Tecnologías de Información

SUB-TÍTULO
Actas de la 7ª Conferencia Ibérica de Sistemas y Tecnologías de Información
Madrid, España
20 al 23 de Junio de 2012

Vol. I – Artículos
Tomo 1

EDITORES
Álvaro Rocha, Universidade Fernando Pessoa
Jose A. Calvo-Manzano, Universidade Politécnica de Madrid
Luís Paulo Reis, Universidade do Minho
Manuel Pérez Cota, Universidad de Vigo

EDICIÓN, IMPRESIÓN Y ACABADO
APPACDM – Associação Portuguesa de Pais e Amigos do Cidadão Deficiente Mental, Braga, Portugal

DEPÓSITO LEGAL
344776/12

ISBN
978-989-96247-6-4

WEB
http://www.aisti.eu/cisti2012