



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria
e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) PI 0717019-0 A2



* B R P I 0 7 1 7 0 1 9 A 2 *

(62) Data de Depósito do Pedido Original:
PI0109685 - 23/03/2001

(22) Data de Depósito: 17/09/2007

(43) Data da Publicação: 08/10/2013
(RPI 2231)

(51) Int.Cl.:

G06N 3/12
G06F 19/00
G06K 9/00

(54) Título: MÉTODO PARA EFETUAR SELEÇÃO DE CARACTERÍSTICA BASEADA EM ALGORITMO GENÉTICO

(30) Prioridade Unionista: 22/09/2006 US 60/826593, 10/01/2007 US 60/884288, 10/01/2007 US 60/884288, 22/09/2006 US 60/826593

(73) Titular(es): Koninklijke Philips Electronics N.V.

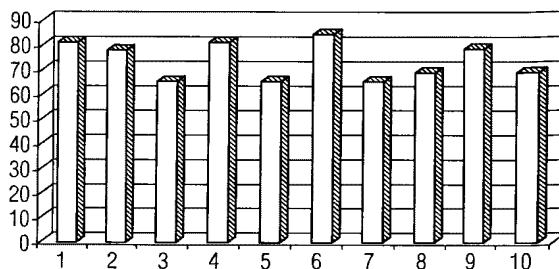
(72) Inventor(es): Lalitha A. Agnihotri, Lilla Boroczky, Luyin Zhao, Michael C. C. Lee

(74) Procurador(es): Momsen, Leonardos & Cia.

(86) Pedido Internacional: PCT IB2007053750 de 17/09/2007

(87) Publicação Internacional: WO 2008/035276de 27/03/2008

(57) Resumo: "SISTEMA E MÉTODO DE FUMO ELÉTRICO". Um sistema de fumo elétrico (21) compreendendo um cigarro (23) e um acendedor elétrico (25), no qual o cigarro (23) compreende uma superfície fosca de tabaco tubular (66) parcialmente cheia com material de tabaco (80) de modo a definir uma porção de barra de tabaco cheia (60) e uma porção de barra de tabaco vaga (90). O cigarro (23) e o acendedor (25) são mutuamente dispostos de modo que quando o cigarro (23) é recebido no acendedor (25), o elemento do aquecedor elétrico (37) do acendedor (25) pelo menos parcialmente sobrepõe pelo menos uma porção da porção da barra de tabaco cheia (60). O cigarro (23) e o acendedor (25) são também mutuamente dispostos de modo que quando o cigarro (23) é recebido no acendedor (25), a extremidade livre (15) do cigarro (23) fica fechada. O cigarro (23) inclui uma zona de perfurações (12,14) em uma localização ao longo da porção da barra de tabaco cheia (60), com o cigarro sendo isento de perfurações ao longo da porção da barra de tabaco vaga (90).



“MÉTODO PARA EFETUAR SELEÇÃO DE CARACTERÍSTICA BASEADA EM ALGORITMO GENÉTICO”

Este pedido reivindica o benefício do pedido provisório da U.S. no. de série 60/826.593 depositado em 22 de setembro de 2006, que é
5 incorporada aqui por inteiro para referência.

Métodos são fornecidos para seleção de característica usando algoritmos genéticos.

Um algoritmo genético (GA) é uma classe de algoritmos evolucionários usados na computação como uma técnica de pesquisa para
10 encontrar soluções para otimização e problemas de pesquisa. GA usa terminologia e conceitos para desenvolver técnicas inspiradas por biologia evolucionária, incluindo conceitos tal como herança, mutação, seleção, e cruzamentos.

Seleção de característica, também conhecida como seleção de
15 subconjunto ou seleção variável, é um método usado em aprendizado de máquina. Antes de aplicar um algoritmo de aprendizado a um conjunto de dados, um subconjunto das características disponíveis do conjunto de dados é selecionado. O processo de seleção de característica é usado porque é computacionalmente infactível usar todas as características disponíveis em
20 um conjunto de dados. Seleção de característica é também usado para minimizar problemas de estimativa e superposição quando um conjunto de dados limitou amostras de dados contendo um grande número de características.

Um domínio típico para o qual a seleção de característica é
25 utilizado, é diagnose auxiliado por computador (CADx). CADx é um método que usa tecnologia de aprendizado de máquina para prognosticar um resultado médico, por exemplo, para classificar lesões desconhecidas como malignas ou benignas. Por exemplo, em formação de imagem por tomografia computadorizada (CT) do pulmão para diagnose de câncer de pulmão, essas

características de entrada podem incluir os resultados de algoritmos de processamento de imagem como aplicado ao nódulo de pulmão sob investigação. Melhorar a precisão de diagnóstico de sistemas de CADx é um passo chave para introdução com sucesso desta tecnologia na clínica.

5 Devido ao grande número de características de imagem e características clínicas que poderiam ser computadas e recuperadas para cada lesão, a seleção de característica é um importante passo devido à inviabilidade de usar todas as características disponíveis em um conjunto de dados e problemas de estimativa quando um conjunto de dados limitou as amostras de
10 dados contendo um grande número de características. Seleção de característica usando GA e máquinas de vetor de suporte (SVM) foi mostrada ser um método eficiente de seleção de característica para detecção auxiliado por computador (CAD; Boroczky et al., IEEE Transaction on Biomedical Engineering, 10(3), pp. 504-551, 2006).

15 Embora a seleção de característica baseada em GA mostrou ser com sucesso em um número de áreas, de problemas e de tendências freqüentemente ocorrem devido aos conjuntos de dados médicos pequenos e ruidosos. Isto é causado por divisão aleatória dentro do GA, que pode gerar conjunto de dados de treinamento tendenciosos e conjunto de dados de teste
20 tendenciosos a partir de um conjunto de dados de aprendizado.

 Consequentemente, métodos são fornecidos aqui para efetuar a seleção de característica baseada em algoritmo genético. Os métodos em uma modalidade incluem os passos para construir múltiplos classificadores para obter pelo menos um resultado de classificação; integrar o pelo menos um
25 resultado de classificação a partir dos múltiplos classificadores para obter uma resultado de precisão integrado; e emitir o resultado de precisão integrado a um resultado de precisão integrado a um algoritmo genético como um valor de aptidão para um subconjunto de características candidatas, onde a seleção de característica baseada em algoritmo genético é efetuada.

Uma modalidade relacionada ainda inclui usar o algoritmo genético para obter o subconjunto de características candidatas.

Em uma modalidade relacionada, os múltiplos padrões de separação de dados dividem o conjunto de dados de aprendizado em dados de
 5 treinamento e dados de teste. Conjuntos de dados de aprendizado são usados para definir parâmetros de uma regra de aprendizado. Um conjunto de dados de treinamento inclui um vetor de entrada (incluindo as características disponíveis) e um vetor de resposta (incluindo diagnose conhecida, i. e. maligno / benigno), e é usado junto com um método de aprendizado os casos
 10 e as diagnoses conhecidas. Um conjunto de dados de teste inclui exemplos conhecidos que são usados para testar o desempenho do classificador construído sobre os dados de treinamento.

Em uma outra modalidade relacionada, os múltiplos classificadores são selecionados a partir de pelo menos um de, uma máquina
 15 de vetor de suporte, uma árvore de decisão, uma análise de discriminante linear, e uma rede neural.

Em uma outra modalidade relacionada, construindo os múltiplos classificadores, ainda inclui usar uma técnica de re-amostragem para obter cada um de uma pluralidade de conjuntos de treinamento e uma
 20 pluralidade de conjuntos de teste a partir do conjunto de dados de aprendizado.

Em ainda uma outra modalidade relacionada, construindo os múltiplos classificadores ainda inclui usar uma pluralidade de conjuntos de treinamento.

25 Em uma outra modalidade, o método ainda inclui combinar resultados de classificação a partir dos múltiplos classificadores para formar um prognóstico de grupo.

Em uma modalidade relacionada, integrar pelo menos um resultado de classificação ainda inclui calcular pelo menos um resultado

selecionado a partir do grupo de uma média, uma média ponderada, um voto majoritário, um voto majoritário ponderado, e um valor mediano.

5 Em uma outra modalidade relacionada, o método ainda inclui usar um algoritmo genético para avaliar repetidamente subconjuntos de características candidatos usando os valores de conveniência, para gerar novos subconjuntos de características candidatos, e obter um subconjunto de características final ótimo.

10 Em uma modalidade relacionada, o método é usado em uma modalidade de formação de imagem médica selecionada do grupo de pelo menos um de CT, MRI, raio X e ultrassom.

Em uma outra modalidade, o método é usado em detecção auxiliada por computador (CAD). Em uma modalidade relacionada, o método é usado em CAD de uma doença selecionada do grupo de pelo menos um de câncer de pulmão, câncer de mama, câncer de próstata, e câncer colorretal.

15 Em ainda uma outra modalidade, o método é usado em diagnose auxiliada por computador (CADx). Em uma modalidade relacionada, o método é usado em CADx de uma doença selecionada do grupo de pelo menos um de câncer de pulmão, câncer de mama, câncer de próstata, e câncer colorretal.

20 Os métodos fornecidos aqui integram métodos de conjunto classificadores em um processo evolucionário de seleção de característica para melhorar a característica seleção de característica baseada em GA. O GA avalia cada subconjunto de característica usando um resultado de prognóstico integrado baseado em múltiplos padrões de separação de dados, mais
25 propriamente do que avaliar um único padrão de separação de dados. Isto é especialmente útil para dados ruidosos que por outro lado podem forçar um cálculo de valor de conveniência tendencioso.

Figura 1 é um gráfico de barras que mostra o impacto da separação dos dados na precisão da classificação.

Figura 2 é um fluxograma que mostra passos da construção de múltiplos classificadores para analisar um conjunto de dados e obter um melhor subconjunto de características.

5 Seleção de característica é usada em determinar um subconjunto de característica ótimo de modo a construir um classificador. Um processo de seleção de características baseado em SVM e GA é usado. Um classificador é construído com base em um subconjunto de características ótimo.

10 Classificadores são usados em CAD e CADx de diferentes doenças, por exemplo para câncer de pulmão e outros tipos de câncer tendo tumores sólidos. No campo de aprendizado de máquina, classificadores são usados para agrupar itens que têm valores de características similares. Possíveis classificadores incluem SVMs, árvores de decisão, análise de discriminante linear e redes neurais. SVMs são classificadores lineares, e são
15 freqüentemente usados já que eles tempo~em mostrado desempenho superior com relação aos classificadores. Uma árvore de decisão é um modelo de prognóstico que mapeia observação sobre um item para conclusões sobre aqueles valores alvos dos itens. Análise de discriminante linear é usada para encontrar a combinação linear de características que melhor separam duas ou
20 mais classes de objetos ou eventos. As combinações resultantes são usadas como um classificador linear ou e, redução de dimensionalidade antes da última classificação. Uma rede neural é uma ferramenta de modelagem de dados estatísticos não lineares que é usada para modelar relações entre entradas e saídas e / ou para encontrar padrões nos dados.

25 Um sistema de CADx que fornece alta confiança para clínicos melhora o andamento do trabalho dos clínicos fornecendo diagnóstico precisos e rápidos (poucos falsos positivos e falsos negativos). Um sistema de CADx pode ser usado como um segundo leitor para aumentar a confiança dos clínicos em seus diagnósticos, levando a uma redução significativa de

biópsias desnecessárias de lesões do pulmão tal com nódulos, e levando a uma redução significativa de atraso de tratamentos desnecessários. Ainda mais, um sistema de CADx pode facilitar o rastreio de câncer de pulmão de pacientes assintomáticos já que o diagnóstico pode ser alcançado rapidamente e com precisão. Digitalizadores MSCT, exemplificados mas não limitados à série Philips Brilliance, oferecem aumento da resolução e permitem estruturas mais finas a serem observados enquanto produzindo quantidades aumentadas de dados de imagem a serem interpretados por radiologistas.

No aprendizado de máquina baseado no domínio da CADx, um da maioria dos problemas comuns é que os dados de treinamento são usualmente ruidosos. Ruído é especialmente prevalente quando um conjunto de dados de treinamento não é suficientemente grande. Isto tem um impacto substancial na eficácia da seleção de característica. Já que GA se baseia na separação dos dados aleatória para avaliar cada cromossomo representando um subconjunto de característica, dados ruidosos fornecem uma avaliação imprecisa de como um conjunto de características atua. Como um resultado, um bom subconjunto de características, pode ser descartado devido ao seu desempenho em uma separação de dados aleatória ‘ ruim ’. Isto, de forma subsequente, afeta a convergência com sucesso para um subconjunto de características ótimo.

Figura 1 mostra um gráfico de resultados de uma experiência usando dados de 129 casos de câncer de pulmão. Um subconjunto de dados selecionados aleatoriamente foi usado para treinamento, i. e. construindo um classificador de SVM, e os dados remanescentes foram usados para teste. Isto é conhecido como separação de dados. O resultado na figura 1 mostra que quando uma diferente separação de dados é usada, a precisão da classificação, i. e. precisão de teste, difere, de forma significativa,.

Métodos anteriores tipicamente assumem que um componente de ruído é aleatoriamente traçado a partir de uma distribuição normal

unbiased, i. e. em torno de zero. O valor de conveniência é tipicamente corrigido estimando uma tendência de ruído e subtraindo-a do valor de conveniência (Miller et al., Evolutionary Computation, 1996, disponível em <http://leitl.org/docs/ecj96.ps.gz>). Um valor de conveniência é uma medida

5 objetiva da qualidade de uma solução.

Nem todos os dados no mundo real tem uma distribuição não tendenciosa, ou uma tendência é difícil de estimar. Para endereçar estes problemas, os métodos fornecidos aqui usam conjuntos classificadores para reduzir o impacto de ruído quando avaliando um subconjunto de característica

10 durante a evolução do GA.

Conjunto classificador tem sido teoricamente e empiricamente provado ser mais preciso do que qualquer dos classificadores individuais realizando os conjuntos (Opitz et al., Journal of Artificial Intelligence Repesquisa, pp. 169-198, 1999). Os métodos fornecidos aqui usam as

15 seguintes variâncias: familiaridade com técnicas de re-amostragem para obter diferentes conjuntos de treinamento para construir múltiplos classificadores, e uso de múltiplos subconjuntos de características para construir múltiplos classificadores. Resultados de classificação a partir de múltiplos classificadores são combinados juntos para formar um prognóstico de grupo.

20 No lugar de construir um classificador (i. e. usando um padrão de separação de dados) de acordo com métodos anteriores para avaliar o desempenho de um subconjunto de características, os métodos fornecidos aqui constroem múltiplos classificadores, também conhecidos como um conjunto, e integram os resultados da classificação a partir desses

25 classificadores. Neste caso, vários classificadores são construídos sobre diferentes separações de dados. Cada classificador vai resultar em uma decisão, e. g. se a lesão é maligna ou benigna. O método de integração pode ser um de voto majoritário, i. e. o prognóstico selecionado pela maioria dos classificadores membros. Métodos de integração alternativos incluem calcular

uma média, uma média ponderada, ou um valor mediano (Kuncheva, L. L, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2), pp. 281-286, 2002). A precisão obtida pelo conjunto/ensemble de classificadores é melhor do que qualquer classificador único. A precisão integrada, como
5 determinada pelo conjunto classificador, é retornada ao GA como um valor de conveniência para um subconjunto específico de características.

Figura 2 mostra amostras de dados separados em dois conjuntos, Conjunto A (o conjunto de dados de aprendizado) e Conjunto B (o conjunto de dados reservado para o teste final). Conjunto A sofre separação
10 dos dados, dividindo os dados do Conjunto A em conjunto de Treinamento e um Conjunto de Teste, Múltiplos padrões de separação de dados são aplicados para construir múltiplos classificadores, i. e. SVM. Os resultados a partir de múltiplos classificadores são integrados e avaliados. Precisão da classificação é efetuada nos dados do conjunto de teste, que é parte do conjunto de dados
15 original. Os resultados da precisão de classificação, que são os resultados integrados a partir de cada classificador, são retornados ao GA como um valor de conveniência para um subconjunto de características candidato. O valor de conveniência pode incluir ambos, especificidade e sensibilidade. Após os resultados integrados serem retornados ao GA, o GA determina que
20 características são retidas / descartadas e gera novo subconjunto(s) de características candidato através de operações de mutação interna e cruzamentos. O processo evolucionário do GA se repete até os critérios de término serem atingidos, quando o melhor subconjunto de característica é determinado.

25 Os métodos fornecidos aqui podem ser usados com várias modalidades de formação de imagem, por exemplo MRI, CT, raio X ou ultrassom. Os métodos fornecidos aqui são aplicados às modalidades de formação de imagem médica, incluindo modalidades de formação de imagem que são usadas para detectar e diagnostica lesões anormais no corpo humano,

por exemplo, dados coletados a partir de sistemas de formação de imagem, i. e. digitalizadores eletrônicos. Os métodos e sistemas fornecidos aqui podem ser usados em estação de trabalho de radiologia, exemplificada mas não limitados à Philips Extended Brilliance Workstation, Philips Mx8000, e a

5 série de digitalizadores Philips Brilliance CT ou incorporada nos sistemas PACS exemplificados mas não limitados à sistemas de Stentor iSite. A invenção fornecida aqui também é usada em CAD e CADx. Quando aplicada à CAD e CADx, a invenção fornecida aqui é usada para detectar e diagnosticar doenças tal como câncer de pulmão, pólipos de cólon, câncer

10 colorretal, câncer de próstata, e câncer de mama e outras lesões cancerosas e não cancerosas.

Ainda mais, será aparente que outras e formas adicionais da invenção, e modalidades outras do que as modalidades específicas e de exemplo descritas acima, podem ser concebidas sem fugir do espírito e

15 escopo das reivindicações anexas e seus equivalentes, e por conseguinte é pretendido que o escopo desta invenção englobe esses equivalentes e que a descrição e reivindicações sejam pretendidas para serem exemplos e não devem ser interpretadas como ainda limitantes. Os conteúdos de todas as referências citadas aqui são incorporados para referência.

REIVINDICAÇÕES

1. Método para efetuar seleção de característica baseada em algoritmo genético, caracterizado pelo fato de compreender:

- aplicar múltiplos padrões de separação de dados a um conjunto de dados de aprendizado para construir múltiplos classificadores para obter pelo menos um resultado de classificação;
- integrar o pelo menos um resultado de classificação a partir de múltiplos classificadores para obter um resultado de precisão integrado; e
- emitir o resultado de precisão integrado para um algoritmo genético como um valor de conveniência para um subconjunto de características candidatas, onde a seleção de característica baseada em algoritmo genético é efetuada.

2. Método de acordo com a reivindicação 1, caracterizado pelo fato de ainda compreender usar o algoritmo genético para obter o subconjunto de características candidatas.

3. Método de acordo com a reivindicação 1, caracterizado pelo fato de que os múltiplos padrões de separação de dados dividem o conjunto de dados de aprendizado em dados de treinamento e dados de teste.

4. Método de acordo com a reivindicação 1, caracterizado pelo fato de que os múltiplos classificadores são selecionados a partir do grupo consistindo de pelo menos um dentre uma máquina de vetor de suporte, uma árvore de decisão, um análise de discriminante linear, e uma rede neural.

5. Método de acordo com a reivindicação 1, caracterizado pelo fato de que construir os múltiplos classificadores ainda compreendem usar uma técnica de re-amostragem para obter cada um de uma pluralidade de conjuntos de treinamento e uma pluralidade de conjuntos de teste a partir dos dados de aprendizado.

6. Método de acordo com a reivindicação 1, caracterizado pelo fato de que construir os múltiplos classificadores ainda compreende usar uma

pluralidade de conjuntos de treinamento.

7. Método de acordo com a reivindicação 1, caracterizado pelo fato de ainda compreender combinar os resultados da classificação a partir dos múltiplos classificadores para forma um prognóstico de grupo.

5 8. Método de acordo com a reivindicação 1, caracterizado pelo fato de que integrar pelo menos um resultado de classificação ainda compreende calcular pelo menos um resultado selecionado a partir do grupo consistindo de uma média, uma media ponderada, um voto majoritário, um voto majoritário ponderado, e um valor mediano.

10 9. Método de acordo com a reivindicação 1, caracterizado pelo fato de ainda compreender usar um algoritmo genético para obter um subconjunto de características finais ótimas.

15 10. Método de acordo com a reivindicação 1, caracterizado pelo fato de que o método é usado em uma modalidade de formação de imagem médica selecionada a partir do grupo consistindo de pelo menos um de, CT, MRI, raio X e ultrassom.

11. Método de acordo com a reivindicação 1, caracterizado pelo fato de que o método é usado em detecção com auxílio do computador (CAD).

20 12. Método de acordo com a reivindicação 11, caracterizado pelo fato de que o método é usado em CAD de uma doença selecionada a partir do grupo consistindo de pelo menos um dentre câncer do pulmão, câncer de mama, câncer de próstata e câncer colorretal.

25 13. Método de acordo com a reivindicação 1, caracterizado pelo fato de que o método é usado em diagnóstico com auxílio de computador (CADx).

14. Método de acordo com a reivindicação 13, caracterizado pelo fato de que o método é usado em CADx de uma doença selecionada a partir de um grupo consistindo de pelo menos um dentre câncer de pulmão, câncer de mama, câncer de próstata e câncer colorretal.

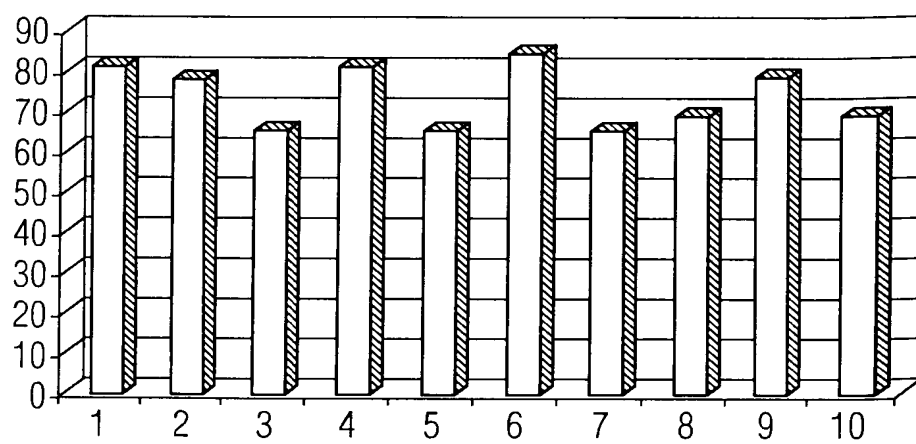


FIG. 1

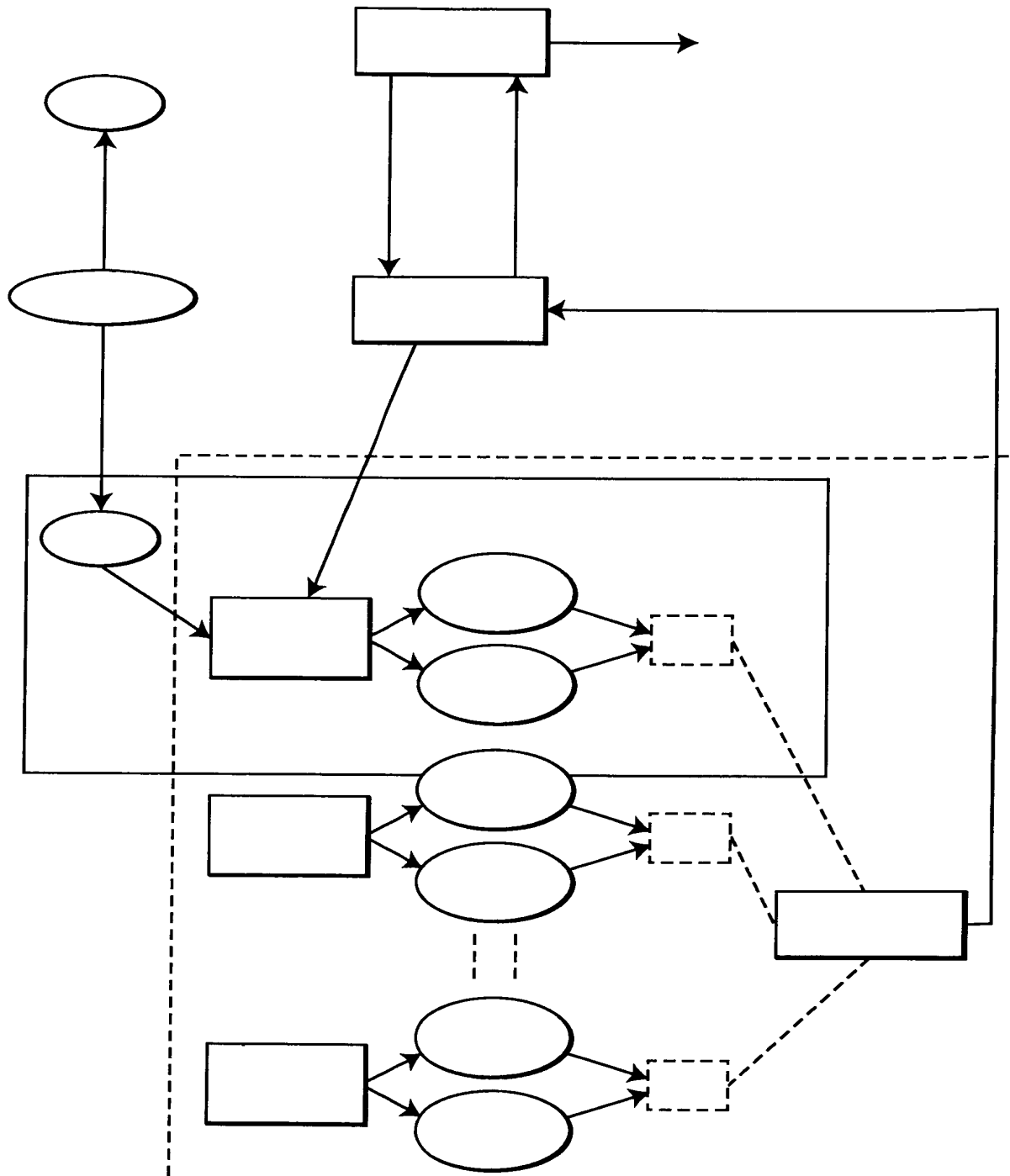


FIG. 2

RESUMO

“MÉTODO PARA EFETUAR SELEÇÃO DE CARACTERÍSTICA BASEADA EM ALGORITMO GENÉTICO”

Métodos para efetuar seleção de característica baseada em algoritmo genético são fornecidos aqui. Em determinadas modalidades, os métodos incluem passos para aplicar múltiplos padrões de separação de dados para um conjunto de dados de aprendizado para construir múltiplos classificadores para obter pelo menos um resultado de classificação; integrar pelo menos um resultado de classificação a partir dos múltiplos classificadores para obter um resultado de precisão integrado; e emitir resultado de precisão integrado para um algoritmo genético como um valor de aptidão para um subconjunto de característica candidata, na qual a seleção de característica baseado em algoritmo genético é efetuada.