SISTEMAS BASEADOS EM ENSEMBLES DE CLASSIFICADORES

INTRODUÇÃO – PROCESSO DE TOMADA DE DECISÕES

- Procuramos uma segunda, terceira ou quarta opinião.
  - Tratando de assuntos financeiros, médicos, sociais entre outros.
- Atribuímos pesos/valores a cada opinião;
- Com a combinação das opiniões se espera obter uma opinião que seja a mais bem informada de todas;
- O processo de consultar "alguns especialistas" antes de tomar uma decisão é um processo da natureza humana.
- Apenas recentemente esse processo foi descoberto pela comunidade de inteligência computacional.

Ensemble based systems (EBS)

- Também conhecido sobe vários outros nomes:
  - Multiple classifier systems, committee of classifiers, ou mixture of experts.
- Tem mostrado produzir resultados favoráveis comparados a sistemas com um único especialista;
- Bons resultados são encontrados em várias aplicações em uma larga variedade de cenários;
- Projeto, implementação e aplicação de tais sistemas são os principais tópicos desta aula.

Razões para Utilizar EBS

- Razões Estatísticas
  - Quando se trabalha com Redes Neurais ou classificadores automatizados:
    - Um bom desempenho no conjunto de treinamento não prediz um bom desempenho de generalização;
    - Um conjunto de classificadores com desempenhos similares no conjunto de classificação podem ter diferentes desempenhos de generalização;
    - Mesmo classificadores com desempenhos de generalização similares podem trabalhar diferentemente;
  - A combinação das saídas produzidas pelos classificadores reduz o risco de uma escolha infeliz por um classificador com um pobre desempenho
    - Não seguir apenas a "recomendação" de um único especialista.

Razões para Utilizar EBS (cont.)

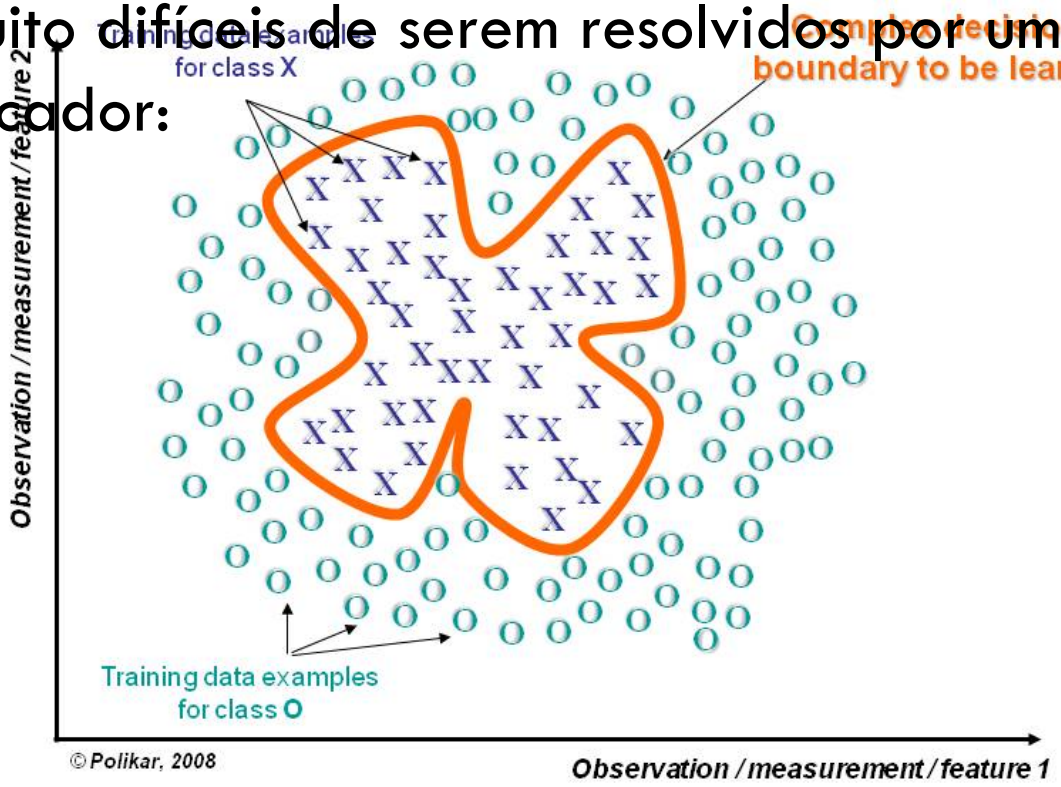
- Grandes volumes de dados
  - A quantidade de dados a serem analisados pode ser muito grande para serem efetivamente manipulados por um único classificador;
    - Análise de transmissão de gás para detecção de vazamento podem gerar 10GB a cada 100km;
    - Projeto similar na UFPE, a cada minuto 2000 vetores com 300 dimensões são coletados;
    - DNA
  - Mais apropriados particionar os dados em sub-conjuntos e treinar diferentes classificadores com diferentes partições dos dados e então combinar as saídas com uma inteligente regra de combinação
    - Geralmente tal estratégia tem se mostrado a mais eficaz.

Razões para Utilizar EBS (cont.)

- Pequenos volumes de dados
  - EBS também podem ser usados diretamente no trabalho em problemas que possuem poucos dados;
  - A disponibilidade de dados para o treinamento de classificadores é de fundamental importância para a obtenção de sucesso;
  - Quando há ausência de dados de treinamento técnicas de re-amostragem podem ser utilizadas para a criação de subconjuntos de dados aleatórios sobrepostos em relação aos dados disponíveis;
    - Cada subconjunto é utilizado para treinar diferentes classificadores e então criar ensembles com desempenhos comprovadamente melhores a modelos solo.

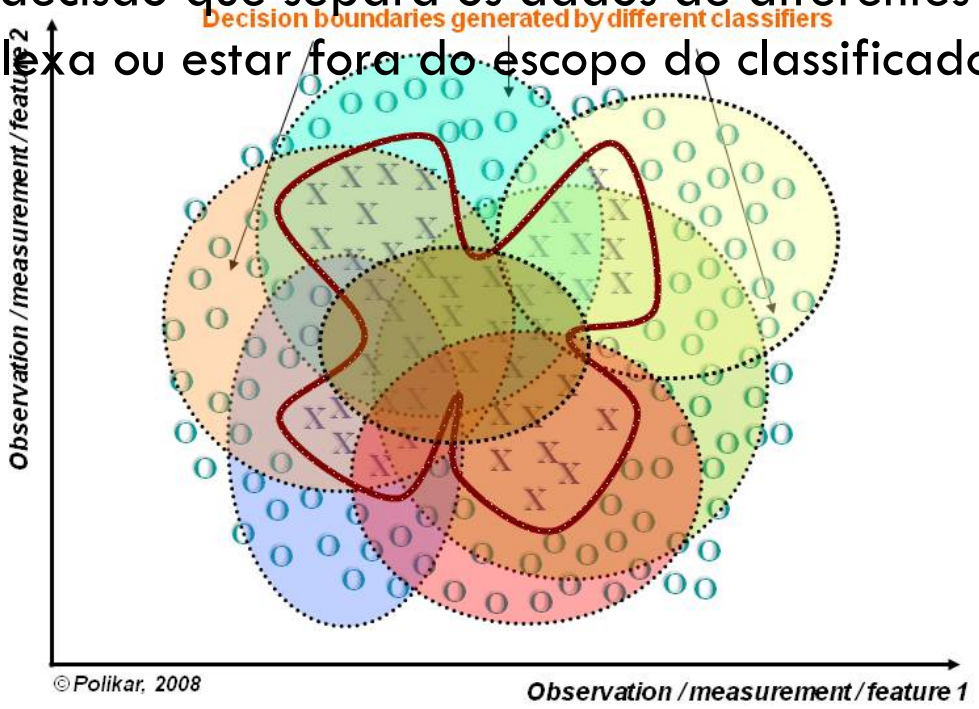
Razões para Utilizar EBS (cont.)

- Dividir e Conquistar
  - Independente da quantidade de dados alguns problemas são muito difíceis de serem resolvidos por um dado classificador:



Razões para Utilizar EBS (cont.)

- Dividir e Conquistar
  - A fronteira de decisão que separa os dados de diferentes classes pode ser muito complexa ou estar fora do escopo do classificador.



Razões para Utilizar EBS (cont.)

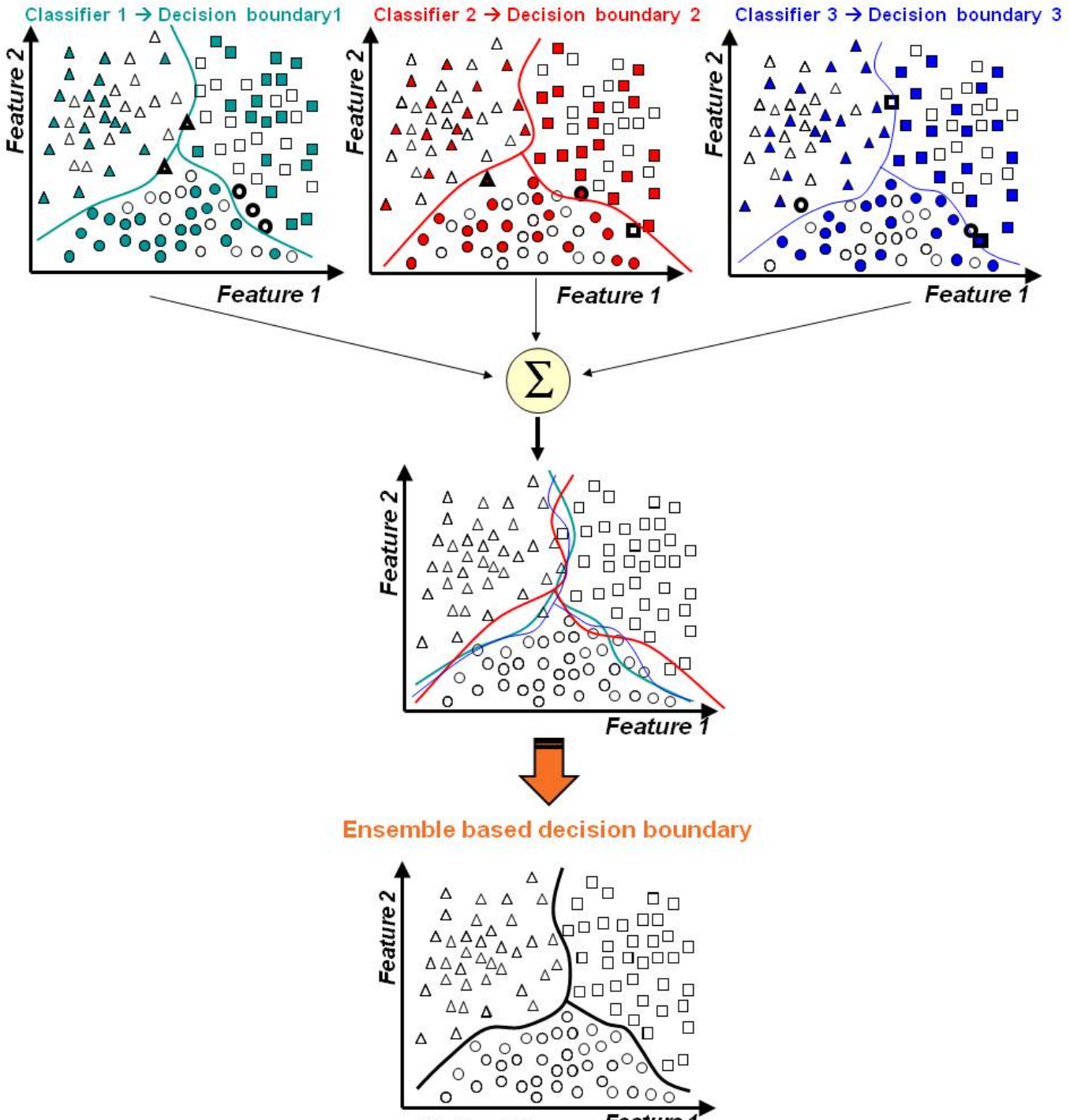
- Dividir e Conquistar
  - A idéia é que o sistema de classificação siga a abordagem dividir-para-conquistar;
  - O espaço de dados é dividido em porções menores e mais "fáceis" de aprender por diferentes classificadores;
  - Assim a linha base da fronteira de decisão pode ser aproximada por meio de uma combinação apropriada dos diferentes classificadores.

Razões para Utilizar EBS (cont.)

- Fusão dos dados
  - A natureza das características/atributos dos dados é heterogênea;
  - Diagnósticos de distúrbios neurológicos: Exames de sangue, Ressonância Magnética, Eletro encefalograma etc.
  - Desconhecimento das fontes e forma de coleta dos dados;
  - Descoberta de características não consideradas na rotulação dos dados do problema.

Razões para Utilizar EBS (cont.)

- Seleção de modelo
  - Considerada a principal razão para o uso de EBS
    - Qual o classificador mais apropriado para um dado problema de classificação?
    - Qual o tipo?: MLP, SVM, Árvores de Decisão, Naive Bayes etc.
    - Qual a configuração?: diferentes inicializações, diferentes amostragens dos dados, etc.
  - Os indivíduos do EBS DEVEM exibir diversidade!


Classifier 1 → Decision boundary1
Classifier 2 → Decision boundary 2
Classifier 3 → Decision boundary 3
Feature 2
Feature 1
Σ
Ensemble based decision boundary

EBS - HISTÓRICO

- Primeiro trabalho datado de 1979 por Dasarathy e Sheela com discussão sobre o particionalmento do espaço de características usando dois ou mais classificadores;
- Em 1990, Hansen e Salamon mostraram que a generalização de uma rede neural pode melhorar usando ensembles;
- Surgimento dos algoritmos de Bagging, Boosting, AdaBoost, novas abordagens, etc.
- Livro sobre Combining Pattern Classifiers: Methods and Algorithms por Ludmila I. Kuncheva em 2004.

EBS - Diversidade

- O sucesso de um EBS, a habilidade em corrigir erros de alguns de seus membros, depende fortemente da diversidade do classificadores que o compõem;
- Cada classificador DEVE fazer diferentes erros em diferentes instâncias dos dados;
- A idéia é construir muitos classificadores e então combinar suas saídas de modo que o desempenho final seja melhor do que o desempenho de um único classificador;
- A diversidade de classificadores pode ser obtida de diferentes formas;

EBS – DIVERSIDADE (CONT.)

- Uso de diferentes conjuntos de dados de treinamentos:
  - Os subconjuntos são normalmente obtidos por meio de técnicas de resampling como bootstrapping ou bagging, na maioria das vezes com reposição.
  - "Classificadores Instáveis" são usados para garantir que as fronteiras geradas pelos indivíduos são adequadamente diferentes, mesmo usando dados de treinamento substancialmente similares;
  - Se os subconjuntos são gerados sem reposição então o processo se chama K-fold;
    - O conjunto de treinamento é dividido em k blocos e cada classificador é treinado em k-1 deles;

EBS – DIVERSIDADE (CONT.)

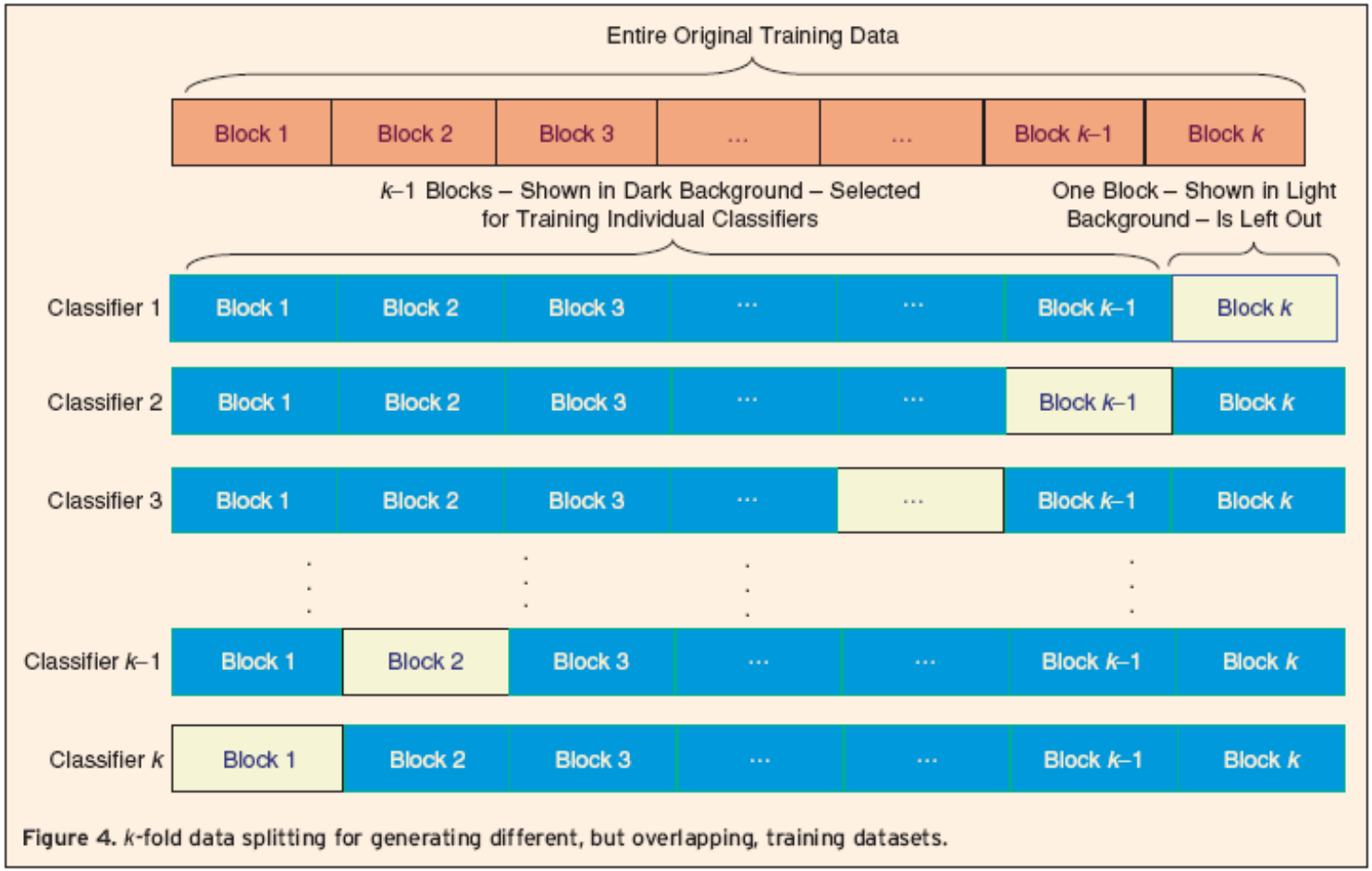


Figure 4. *k*-fold data splitting for generating different, but overlapping, training datasets.

EBS – DIVERSIDADE (CONT.)

- Outra abordagem para se obter diversidade é o uso de diferentes parâmetros de treinamento para diferentes classificadores:
  - Redes Neurais
    - Usando diferentes conjuntos de pesos iniciais; numero de camadas/nodos; funções de ativação; algoritmos de treinamento e seus parâmetros.
- Usar diferentes tipos de classificadores;
- Usar diferentes conjuntos de características;
- A forma mais tratável, usada e recomendada para inserir diversidade em um EBS é através da manipulação do conjunto de treinamento.

EBS – MEDIDAS DE DIVERSIDADE

- Existem propostas para avaliar quantitativamente a diversidades dos classificadores;
- Uma das mais simples é a medida por pares. Para T classificadores podemos calcular T(T-1)/2 medidas de diversidade pareadas, então a diversidade total do ensemble pode ser obtida pela média dos pares;
- Dada duas hipóteses Hi e Hj:

- Onde

	Hj é correto	Hj é incorreto
Hi é correto	a	c
Hi é incorreto	b	d

EBS – MEDIDAS DE DIVERSIDADE (CONT.)

- Correlação: a diversidade é medida como a correlação entre as saídas de dois classificadores

$$\rho_{i,j} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \quad 0 \leq \rho \leq 1.$$

- Q-Statistic
  - Qij = (ad – bc)/(ad + bc)
  - Q assume valores positivos se as instâncias são corretamente classificadas por ambos os classificadores e valores negativos caso contrário;
  - Assim como na correlação uma alta diversidade é obtida com 0.

EBS – MEDIDAS DE DIVERSIDADE (CONT.)

- Medidas de desacordo e falta dupla
  - Dij = b + c
  - DFij = d.
- Entropia

$$E = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T - \lceil T/2 \rceil} \min \{\zeta_i, (T - \zeta_i)\}$$

  - N é a cardinalidade da base, T número de classificadores; [.] operador ceiling e zeta o no. de classif. que classificam incorretamente um padrao **X**i
  - A entropia varia entre 0 e 1: 0 indica que todos os classificadores são praticamente os mesmos e 1 indica uma alta diversidade
- Variância de Kohavi-Wolpert similar a Dij.

EBS – DOIS COMPONENTES CHAVE

- Escolha da estratégia para a construção de um EBS composto por classificadores o mais diverso quanto possível:
  - Algumas estratégias são: Bagging, Boosting, AdaBoost, Stacked Generalization e Mixture of Experts.

- Estratégia necessária para a combinação das saídas de cada classificador que compõem o EBS:
  - Combinação que deve amplificar a quantidade de decisões corretas e anular as ocorrência das incorretas.

Criando um Ensemble

- Como os classificadores serão gerados?
- Como tais classificadores irão diferir entre eles?
- Respostas -> determinarão a diversidade dos classificadores = performance final do EBS;
- Uma estratégia para geração dos membros de um EBS DEVE buscar uma melhora da diversidade;
- Não existe uma única medida de diversidade aceita uniformemente;
- O aumento da diversidade em EBS é tratado com um problema de busca - com emprego de heurísticas - usando procedimentos de resampling ou seleção de diferentes parâmetros de treinamento.

Algoritmo Bagging

- O primeiro algoritmo para a construção de EBS;
- Possui uma implementação simples e intuitiva;
- A diversidade é obtida com o uso de diferentes subconjuntos de dados aleatoriamente criados com reposição;
- Cada subconjunto é usado para treinar um classificador do mesmo tipo;
- As saídas dos classificadores são combinadas por meio do voto majoritário com base em suas decisões;
- Para uma dada instância, a classe que obtiver o maior número de votos será então a resposta.

Algoritmo Bagging

Algorithm: Bagging

Input:

- Training data S with correct labels $\omega_i \Omega = \{\omega_1,...,\omega_C\}$ representing C classes
- Weak learning algorithm **WeakLearn,**
- Integer T specifying number of iterations.
- Percent (or fraction) F to create bootstrapped training data

Do t=1, ..., T

1. Take a bootstrapped replica S_t by randomly drawing F percent of S.
2. Call **WeakLearn** with S_t and receive the hypothesis (classifier) h_t.
3. Add h_t to the ensemble, $\mathcal{E}$.

End

Test: Simple Majority Voting – Given unlabeled instance **x**

1. Evaluate the ensemble $\mathcal{E} = \{h_1,...,h_T\}$ on **x**.
2. Let $v_{t,j} = \begin{cases} 1, & \text{if } h_t \text{ picks class } \omega_j \\ 0, & \text{otherwise} \end{cases}$ be the vote given to class ω_j by classifier h_t.
3. Obtain total vote received by each class , $V_j = \sum_{t=1}^{T} v_{t,j}$ j = 1,...,C.
4. Choose the class that receives the highest total vote as the final classification.

ALGORITMO BAGGING - VARIAÇÕES

- Random Forests
  - Usado para a construção de EBS com árvores de decisão;
  - Variação da quantidade de dados e características;
  - Usando árvores de decisão com diferentes inicializações;
- Pasting Small Votes
  - Segue a idéia do bagging, mas voltado para grande volumes de dados;
  - A base de dados é dividida em subconjuntos chamados de *bites*;
  - Trabalha com as instâncias mais informadas.

Algoritmo Boosting

- Criado em 1990 por Schapire é considerado o mais importante desenvolvimento na história recente da aprendizagem de máquina;
- Também cria EBS por meio da re-amostragem dos dados;
- A re-amostragem é estrategicamente criada para prover o conjunto de treinamento mais informativo para cada classificador;
- Normalmente o EBS possui apenas três classificadores;
- Comprovadamente a performance do EBS é melhor que a performance do melhor indivíduo.

Algoritmo Boosting

Algorithm: Boosting

Input:

- Training data S of size N with correct labels ω_i, $\Omega = \{\omega_1, \omega_2\}$;
- Weak learning algorithm **WeakLearn**.

Training

1. Select $N_1 < N$ patterns without replacement from S to create data subset S_1.
2. Call **WeakLearn** and train with S_1 to create classifier C_1.
3. Create dataset S_2 as the most informative dataset, given C_1, such that half of S_2 is correctly classified by C_1, and the other half is misclassified.:
   a. Flip a fair coin. If Head, select samples from S, and present them to C_1 until the first instance is misclassified. Add this instance to S_2.
   b. If Tail, select samples from S, and present them to C_1 until the first one is correctly classified. Add this instance to S_2.
   c. Continue flipping coins until no more patterns can be added to S_2.
4. Train the second classifier C_2 with S_2.
5. Create S_3 by selecting those instances for which C_1 and C_2 disagree. Train the third classifier C_3 with S_3.

Test – Given a test instance **x**

1. Classify **x** by C_1 and C_2. If they agree on the class, this class is the final classification.
2. If they disagree, choose the class predicted by C_3 as the final classification.

Algoritmo AdaBoost

- O Adaptive Boosting foi criado por Freund and Schapire em 1997;
- É uma versão mais genérica do algoritmo de boosting original;
- Foram criados os AdaBoost.M1 e AdaBoost.R para manipulação de múltiplas classes e para problemas de regressão, respectivamente;
- O AdaBoost gera um conjunto de hipóteses e as combina por meio da votação ponderada;
- As hipóteses são geradas por meio do treinamento de classificadores usando uma distribuição dos dados iterativamente ajustada.

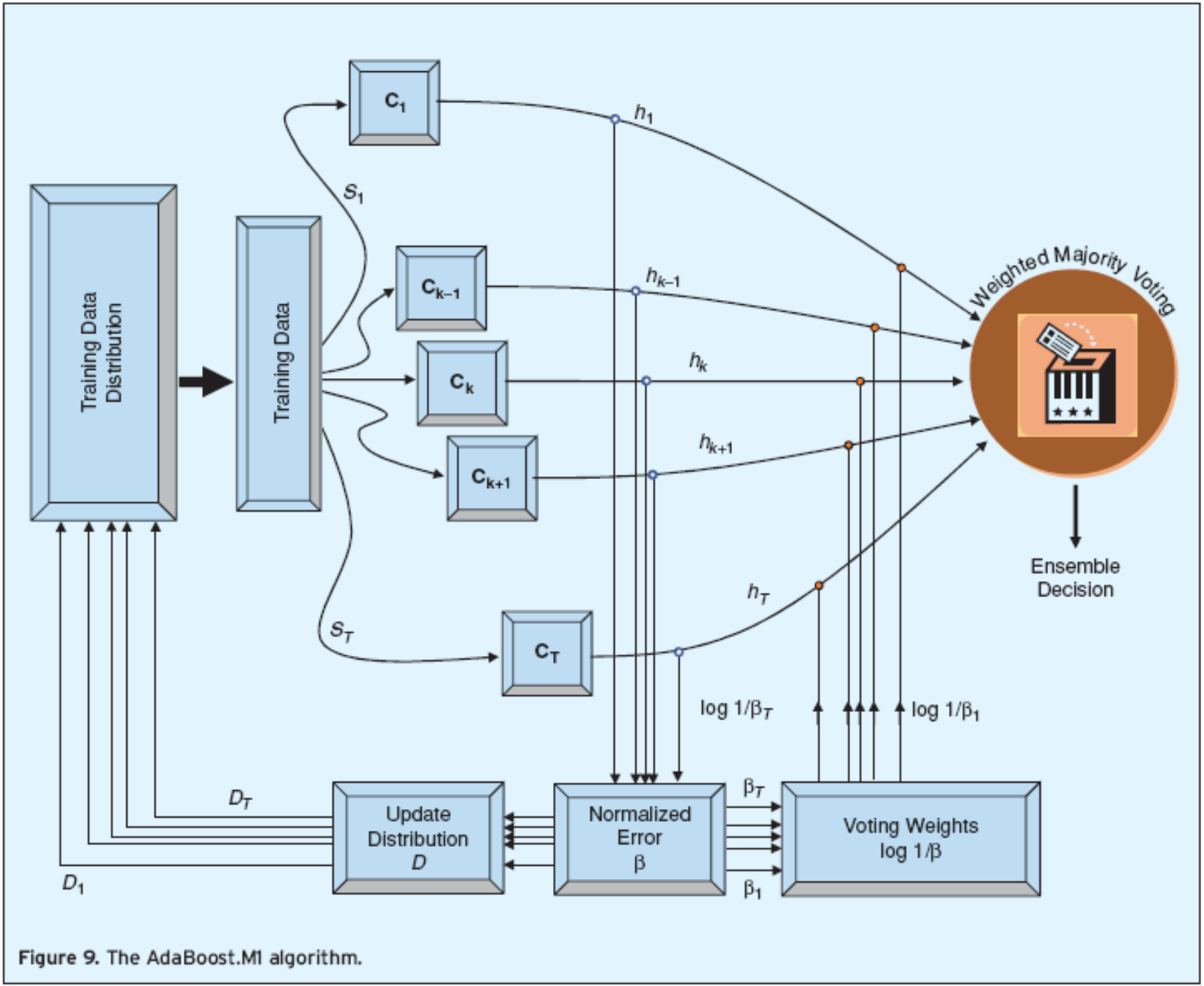


Figure 9. The AdaBoost.M1 algorithm.

Stacked Generalization

- Como aprender a forma de erro e acerto dos classificadores?
- Como mapear as saídas dos classificadores em relação as saídas verdadeiras?
- Os classificadores do EBS são criados usando k-fold, por exemplo;
- As saídas desses classificadores são usadas como entrada para um meta-classificador com o objetivo de aprender o mapeamento entre as saídas e as classes corretas;
- Após o treinamento do meta-classificador os classificadores primários são re-treinados.

Stacked Generalization

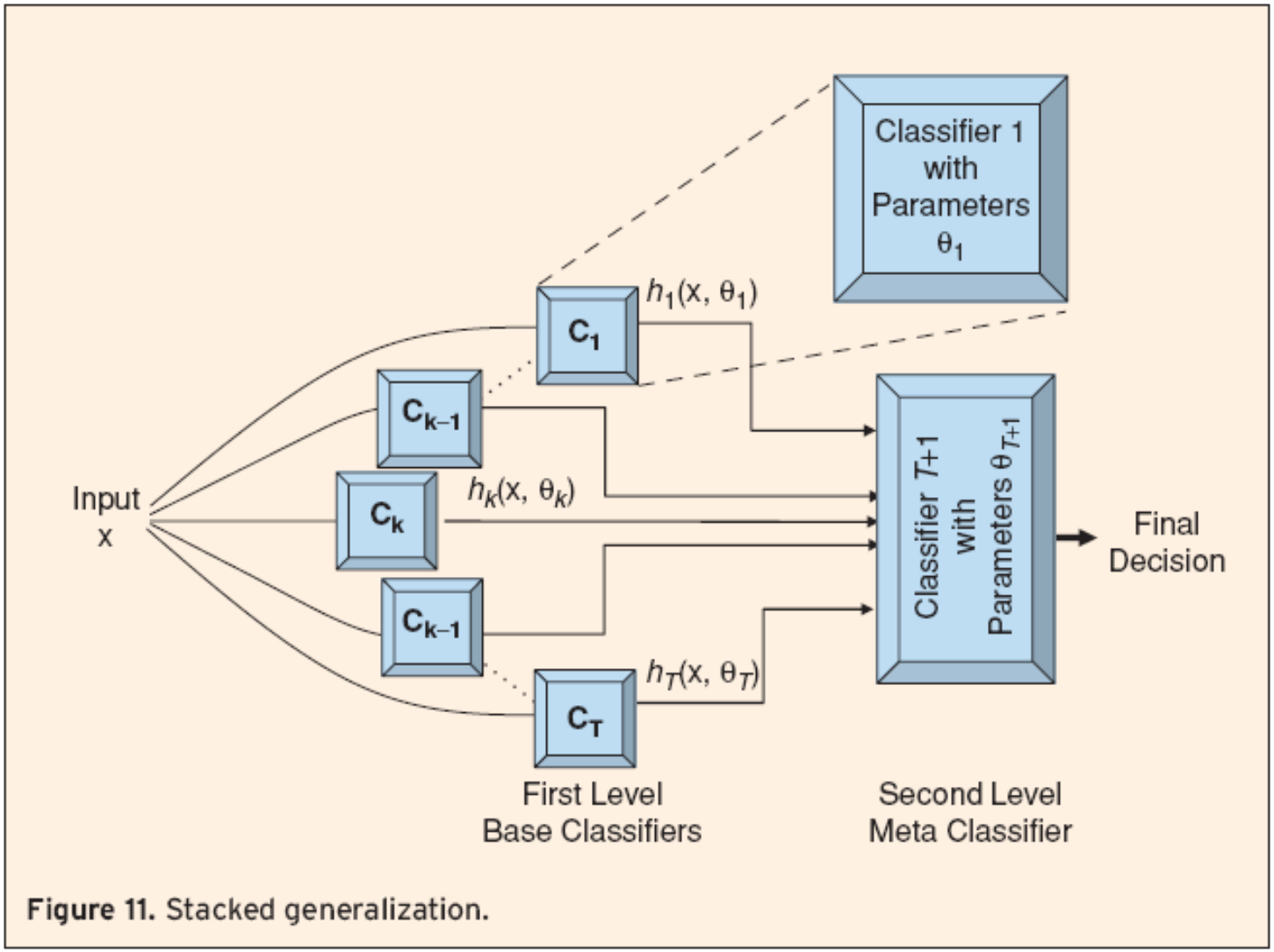


Figure 11. Stacked generalization.

Mixture of experts

- Similar ao Stacked Generalization aonde existe um classificador extra ou meta-classificador;
- Neste caso o classificador no segundo nível é usado para atribuir pesos aos classificadores;
- Atualiza a distribuição dos pesos que é utilizada pelo módulo de combinação das decisões;
- O classificador secundário normalmente é uma *gating networks* treinada com gradiente descendente ou Expectation Maximization (EM);
- Tem-se uma regra de combinação dinâmica;
- Os classificadores devem gerar saídas em valores contínuos.

Mixture of experts

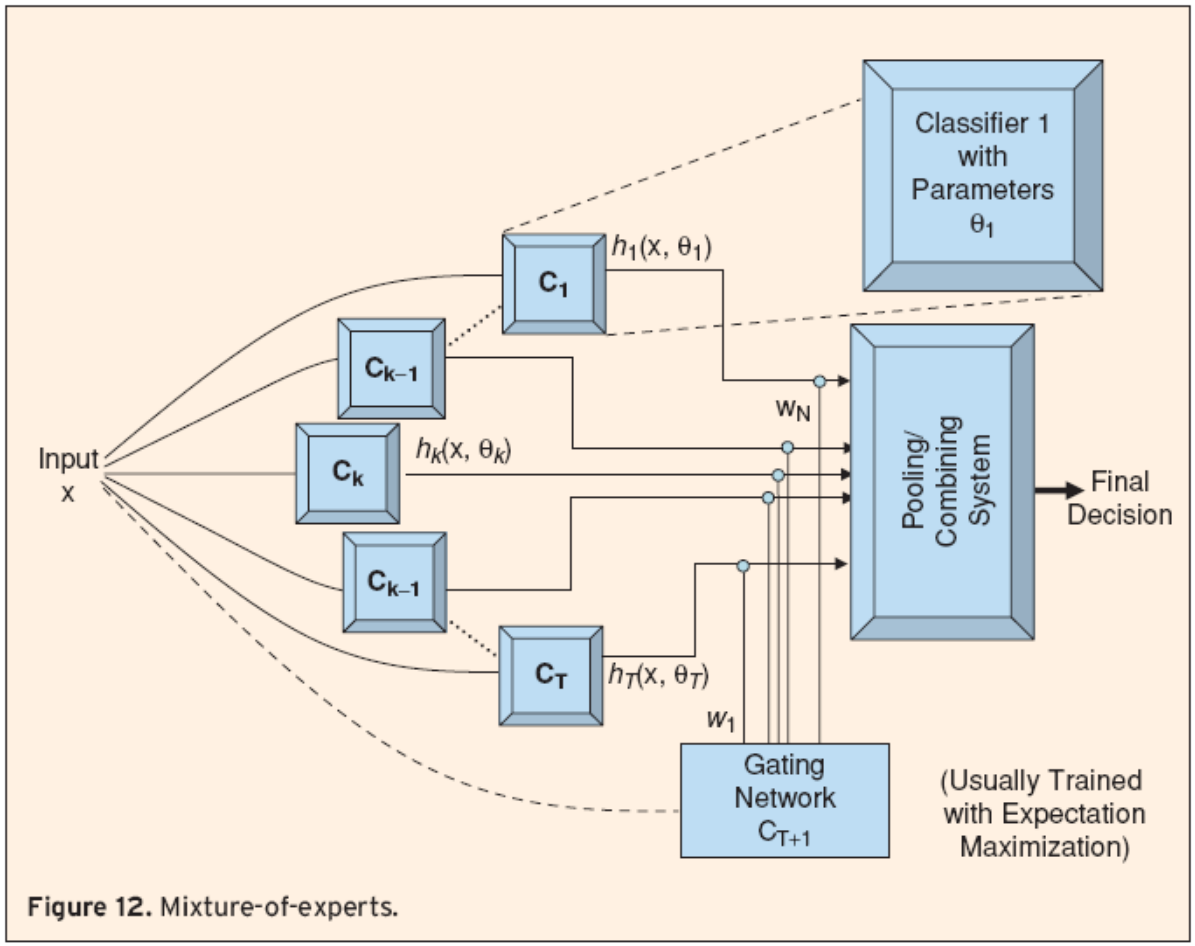
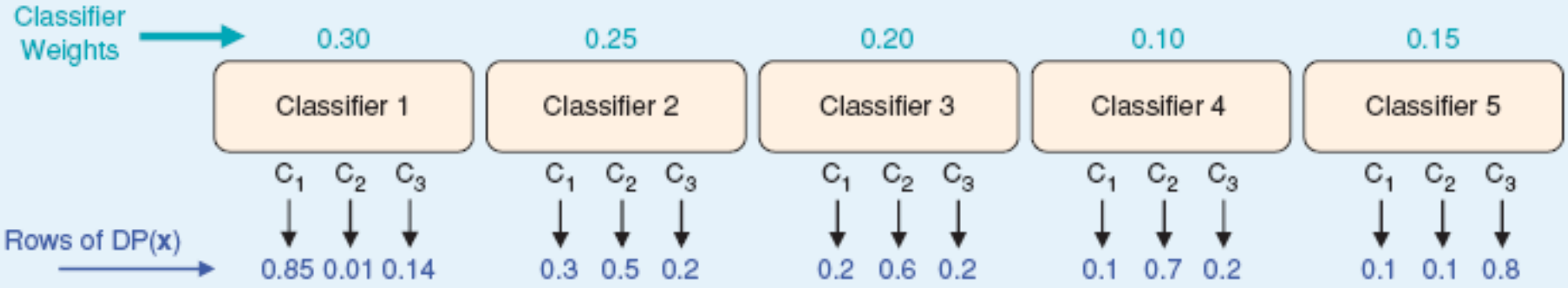


Figure 12. Mixture-of-experts.

MÉTODOS DE COMBINAÇÃO

- Métodos Algébricos
  - Média
  - Média ponderada
  - Soma
  - Soma ponderada
  - Produto
  - Máximo
  - Mínimo
  - Mediana
- Métodos baseados em votação
  - Votação Majoritária
  - Votação Majoritária Ponderada
  - Borda count



Product Rule:	**C_1: 0.0179,**	C_2: 0.00021,	C_3: 0.0009
Sum Rule:	C_1: 1.55,	**C_2: 1.91,**	C_3: 1.54
Mean Rule:	C_1: 0.310,	**C_2: 0.382,**	C_3: 0.308
Max Rule:	**C_1: 0.85,**	C_2: 0.7,	C_3: 0.8
Median Rule:	C_1: 0.2,	**C_2: 0.5,**	C_3: 0.2
Minimum Rule:	C_1: 0.1,	C_2: 0.01,	**C_3: 0.14**
Weighted Average	**C_1: 0.395,**	C_2: 0.333,	C_3: 0.272
Majority Voting:	C_1: 1,	**C_2: 3,**	C_3: 1 Vote
Weighted Majority Voting:	C_1: 0.3,	**C_2: 0.55,**	C_3: 0.15 Votes
Borda Count:	C_1: 5	**C_2: 6,**	C_3: 4 Votes

Figure 15. Example on various combination rules.

Random Forests

- Ensemble method specifically designed for decision tree classifiers
- Random Forests grows many classification trees
- Ensemble of unpruned decision trees
- Each base classifier classifies a "new" vector
- Forest chooses the classification having the most votes (over all the trees in the forest)

Random Forests

- Utiliza dois tipos de aleatoriedade: "Bagging" e "Random input vectors"
  - Cada árvore é gerada usando amostras bootstrap do conjunto de treinamento
  - Em cada nó, o melhor split é escolhido de uma amostra de m_{try} atributos, ao invés de todos os atributos

Random Forests

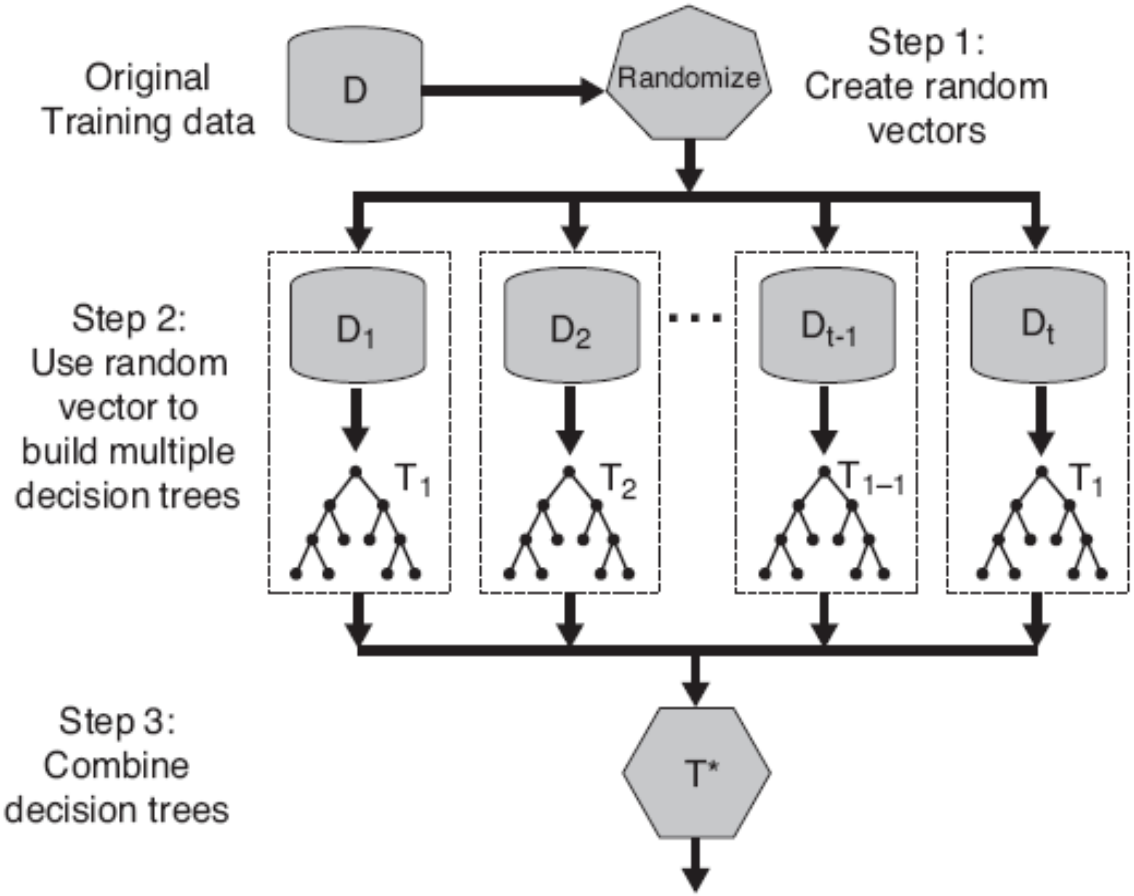


Figure 5.40. Random forests.

Random Forest Algorithm

- M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node.
- m is held constant during the forest growing
- Each tree is grown to the largest extent possible
- There is no pruning
- Bagging using decision trees is a special case of random forests when m=M.

Random Forest Algorithm

- Out-of-bag (OOB) error
- Boa precisão sem over-fitting
- Algoritmo rápido; facilmente paralelizável
- Trata dados de alta dimensionalidade sem maiores problemas
- Only one tuning parameter $m_{try} = \sqrt{p}$, usually not sensitive to it

EBS - Questões

- Qual a melhor estratégia para inserção de diversidade?
  - Alguns estudos o método de boosting se mostra melhor na média, porém é muito sensível a ruídos e outliers;
- Qual o melhor método de combinação das decisões?
  - Totalmente dependente do problema a ser resolvido;
  - Muitos preferente o uso da média, devido a sua simplicidade e desempenho consistente;
  - Método baseados em votação são cada vez mais encontrados em trabalhos recentes.

No Free Lunch Theorem!

EBS – ÁREAS EMERGENTES

- Aprendizagem incremental
  - Aprender sem esquecer o que foi aprendido
- Fusão de dados
  - Trabalhar com diferentes fontes de dados
- Seleção de características
  - Encontrar a quantidade de características apropriadas
- Error Correcting Outputs Codes
  - Decomposição de problemas multiclasses
- Confidence Estimation
  - O quão segura é a decisão tomada pelo EBS
- Uso em ambientes dinâmicos e para aprendizagem não supervisionada

REFERÊNCIA BIBLIOGRÁFICAS

- S. Haykin, Neural Networks: A Comprehensive Foundation. Prentice Hall, 1999.
- R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine, vol. 6, no. 3, pp. 21–45, Quarter 2006.*
- L. Kuncheva, Combining pattern classifiers: methods and algorithms. Wiley-Interscience, 2004.
- T. Dietterich, "Ensemble methods in machine learning," in *Proceedings of the First International Workshop on Multiple Classifier Systems*. London, UK: Springer-Verlag, 2000, pp. 1–15.