# MINIMIZANDO OS RISCOS EM ESTUDOS DE SEGMENTAÇÃO UTILIZANDO CLUSTER ENSEMBLES

## *MINIMIZING RISKS IN SEGMENTATION STUDIES USING CLUSTER ENSEMBLES*

■ **LUIZ CAMPOS DE SÁ LUCAS**

GRADUADO EM ENGENHARIA PELA **PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO (PUC-RJ);** MESTRE EM PROGRAMAÇÃO MATEMÁTICA PELA **UNIVERSIDADE FEDERAL DO RIO DE JANEIRO (UFRJ-COPPE).** É DIRETOR TÉCNICO DA **IDS MARKET INTELLIGENCE** E O ATUAL REPRESENTANTE NO BRASIL DA **ESOMAR.**

E-MAIL: LUIZSALUCAS@IDSBR.COM.BR

**RESUMO**

Recentemente surgiu uma técnica genérica de aprendizado mecânico que se mostrou bastante promissora: os conjuntos de aprendizado. No caso de classificadores, um bom exemplo é o caso da técnica de Random Forests, que apresenta, entre outras, características de excelente capacidade de generalização e bom comportamento em amostras pequenas. A idéia geral é produzir múltiplas soluções e basear-se em algum mecanismo que permita obter uma "solução de consenso" entre essas alternativas. É possível aplicar essa idéia no caso de *cluster analysis,* uma etapa crucial em estudos de segmentação em pesquisas de marketing. Nesse caso a denominação constante da literatura sobre o assunto é *Cluster Ensembles*. Criando várias soluções diferentes, é possível obter uma solução de consenso que apresente várias características interessantes, como será apresentado neste artigo. O foco aqui é a minimização do risco decorrente de uma má escolha do método de *clusterização.*

**PALAVRAS-CHAVE:**

*Cluster ensembles, cluster analysis,* estudos de segmentação.

*ABSTRACT*

*Learning techniques as a general approach for machine learning have shown a great recent promise. In the classifiers area, for example, a good example is Random Forests, a technique that shows a great generalization power, robustness for small samples etc. The general idea is to produce multiple solutions and, based on some mechanism, produce a "consensus solution" among these alternatives. It is possible to apply this idea to cluster analysis, a critical step in segmentation studies in marketing research. In this area this approach is named in the literature as Cluster Ensembles. Producing several different solutions, it is possible to get a consensus solution that presents several interesting features, as we show in the paper. Here the focus is on the minimization of de risk caused by a bad choice of the clustering algorithm.*

*KEY WORDS:*

*Cluster ensembles, cluster analysis, studies of segmentation.*

## 1. INTRODUÇÃO

Na área de aprendizado supervisionado, como por exemplo, em classificadores *(classifiers),* várias técnicas têm se mostrado úteis, tais como:

- *Boosting:* utilização de um conjunto de técnicas menos precisas *(weak learners)* para obter um resultado de consenso que tenha boa qualidade, já que uma solução complementa a outra.
- *Bagging:* uso de uma pluralidade de votos (ou mesmo de amostras) como base para uma boa solução, que seja generalizável.

Um bom exemplo disso é a técnica de Random Forests (BREIMAN, 2001), que apresenta excelentes características de precisão, generalização para outras amostras que não aquela em que o classificador foi treinado e capacidade de bom desempenho em pequenas amostras. Para uma aplicação da técnica na determinação da Importância Derivada em, por exemplo, estudos de satisfação em pesquisas de mercado, veja Soares e Esteves (2008).

O que aqui se apresenta, é como utilizar uma técnica semelhante (geração de várias soluções seguida da criação de uma solução de consenso entre essas alternativas), que apresente características desejáveis. Essa técnica é usualmente descrita na literatura como *Cluster Ensembles* (veja, por exemplo, Hornik (2007), Retzer e Shan (2007) e Orme e Johnson (2008)).

O foco é a minimização do risco decorrente de uma má escolha de um algoritmo de *clusterização*. Sabe-se que diferentes algoritmos de *cluster analysis,* podem produzir soluções completamente diferentes, dependendo das características do método e da estrutura da amostra referente ao problema. Esses aspectos serão descritos a seguir mais detalhadamente.

O método é bastante próximo daquele descrito em Orme e Johnson (2008), embora tenha características de uma aplicação mais ampla. A idéia é aplicar diferentes métodos ao problema e reaplicar, numa segunda etapa, uma análise de *clustering* ao conjunto de agrupamentos definido na etapa anterior. Foram analisados oito diferentes métodos aplicados a catorze bases de dados diferentes, com diferentes graus de complexidade do problema.

## 2. DESCRIÇÃO DO MÉTODO

Pode-se adotar a definição dada por Retzer e Shan (2007): considere $P_1, P_2, \ldots, P_L$ um conjunto de partições de uma base de dados Z. O objetivo é determinar P*, baseando-se em $P_1$-$P_L$, que melhor represente a estrutura de Z. P* é a solução combinada que corresponde a um consenso. A escolha dos algoritmos cabe ao analista: o que importa é que haja suficiente diversidade entre as soluções individuais.

Retzer e Shan (2007) e Orme e Johnson (2008) indicam várias maneiras de se definir as soluções individuais:

- No caso de algoritmos como o k-médias (e outros, como será visto a seguir), podem-se gerar várias soluções iniciais de forma aleatória.
- *Bootstraping*/sub-amostragem e assemelhados, variando-se assim as bases amostrais (MINAEI-BIDGOLI et al., 2004).
- Utilização de vários tipos de algoritmos (como é o caso deste artigo).
- Utilização de subconjuntos das variáveis de segmentação.
- Escolha/variação aleatória do número de *clusters*/grupos.
- Projeção dos dados em subespaços afins ou em subconjuntos dos Componentes Principais (neste caso, escolha aleatória, já que a escolha dos componentes principais mais explicativos do ponto de vista da variância/correlação no caso geral não funciona bem) Yeung e Ruzzo (2001).

Para fins deste estudo optou-se por utilizar os seguintes métodos de *cluster analysis:*

- Classes latentes (flx).
- K-médias (kmn).
- Algoritmos hierarquizados:
  - Divisivo (div).
  - Ligações simples *(single linkage* — sng).
  - *Ward* (wrd).
- *Fuzzy clustering* (fny).
- K-medianas (kmd).
- *Simulated Annealing* (san).

Os algoritmos de classes latentes, k-médias, hierarquizados e de *fuzzy clustering* estão detalhadamente descritos em Wedel e Kamakura (2000).

Os algoritmos de classes latentes são particularmente interessantes num problema de segmentação onde exista uma variável dependente, porém este não é o caso. Uma dificuldade adicional é que o método pode apresentar problemas de convergência quando existe apenas uma observação por respondente, como geralmente é o caso em estudos de segmentação. Em diversas bases de dados testadas neste estudo, o algoritmo não convergiu.

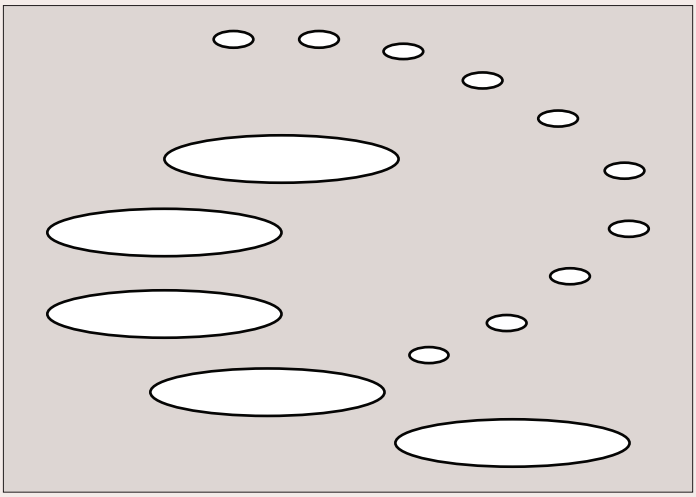

**FIGURA 1**
Estrutura típica de grupos alongados.

O algoritmo k-médias é talvez o algoritmo mais popular em segmentação. Como normalmente a solução inicial do método é aleatória, é comum se resolver o problema várias vezes, adotando a melhor entre as soluções geradas. Neste caso foram geradas dez soluções iniciais em cada base de dados.

O método hierarquizado de Ligação simples tem a característica de identificar grupos mais alongados, como exemplificado na Figura 1.

O método da Ligação simples é o mais capaz de identificar dois grupos na Figura 1: as elipses maiores e as elipses menores.

O método de *Ward* tende a formar grupos compactos e a ter um excelente desempenho em segmentação. Já o método Divisivo foi o que apresentou o pior desempenho nas bases de dados estudadas.

O método *Fuzzy* teve excelente desempenho, e tem a característica extremamente interessante de definir, para cada observação, um coeficiente de pertinência em relação a cada grupo, obtendo-se assim uma probabilidade de pertinência da observação a cada grupo. Obviamente uma boa segmentação deve apresentar, para cada observação da amostra, uma alta probabilidade de pertinência para um dos grupos, com probabilidades baixas para os demais.

O método K-medianas tem uma estrutura semelhante ao método K-médias. A diferença é que a semente/centro do grupo é uma mediana e não uma centróide (média). A mediana de um grupo é o ponto que minimiza a soma das distâncias aos demais elementos do mesmo grupo, conforme ilustra a Figura 2.

Dentre os cinco pontos da figura, o ressaltado em preto é

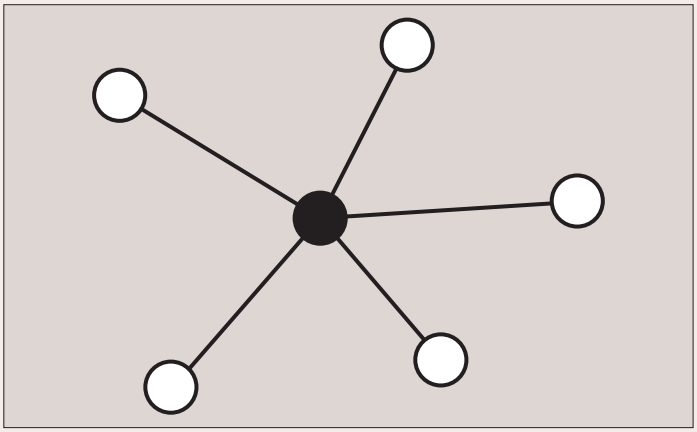

**FIGURA 2**
Mediana de um grupo.

aquele cujas distâncias aos demais elementos é mínima. Um método baseado em medianas (também denominadas medóides) é, como no caso geral de medianas, mais robusto, menos sujeito a efeitos indesejáveis decorrentes de *outliers*.

Assim o método pode ser visto no caso de uma segmentação em *m* grupos, como uma seleção de *m* medianas que minimize a soma das distâncias dos demais elementos da amostra a esse conjunto de medóides. Cada elemento é então alocado ao grupo cuja mediana seja mais próxima.

Outra característica interessante do método K-medianas é que ele pode ser inicializado com um algoritmo "guloso":

- Calcula-se inicialmente a mediana de toda a amostra.
- Efetua-se uma segmentação em dois grupos, mantendo a mediana anterior e selecionando uma segunda mediana de tal forma que a soma dos elementos à mediana de seu grupo, seja mínima.
- Em cada etapa subseqüente adiciona-se uma nova mediana ao conjunto, até que se obtenha *m* grupos.

Como aqui proposto, ou seja, o K-medianas com inicialização pelo algoritmo guloso tem a enorme vantagem de sempre gerar a mesma solução para uma mesma amostra, o que torna o modelo muito estável.

Da mesma forma que o K-médias, o K-medianas não garante um ótimo global. Na verdade todos os métodos aqui analisados se constituem em heurísticas, ou no caso do *Simulated Annealing*, numa meta-heurística, e em nenhum dos casos pode-se garantir a obtenção de um ótimo global.

*Simulated Annealing* é um método geral de otimização e está detalhadamente descrito em Henderson et al. (2003). Na aplicação a que se refere esse artigo, trata-se de escolher um conjunto de *m* medóides que minimize a soma das distâncias dos elementos aos seus centros de grupo (medianas). A função objetivo então é a mesma do algoritmo k-medianas.

Uma vez obtida uma solução em cada um dos oito métodos anteriores, passa-se a um novo problema: em cada observação, tem-se o grupo a que essa observação pertence. Assim tem-se uma nova base de dados com, neste caso, oito variáveis nominais. Pode-se então utilizar uma função de distância que possa ser aplicada a variáveis nominais (como o coeficiente generalizado de Gower — veja Wedel e Kamakura (2000)) e aplicar um método que trabalhe com distâncias. Nesse exemplo foi adotado o algoritmo k-medianas.

## 3. BASES DE DADOS UTILIZADAS

Para o teste, utilizaram-se catorze bases de dados, constantes da Tabela 1:

**TABELA 1**
Bases de dados.

| DATASET | FONTE | GRUPOS REAIS EXISTENTES | NÚMERO DE VARIÁVEIS |
|---|---|---|---|
| Íris | Weka/Uci | 3 | 4 |
| Cassini | Clue | 3 | 2 |
| Dados1 | Qiu e Joe | 7 | 8 |
| Dados2 | Qiu e Joe | 5 | 8 |
| Cpu | Weka | 7 | 19 |
| *Segment-challenge* | Weka | 7 | 32 |
| *Breast-cancer* | Uci | 2 | 6 |
| *Segment-test* | Weka | 4 | 19 |
| *Ecoli* | Uci | 8 | 7 |
| *Cloud1* | Uci | 5 | 9 |
| *Mammographics* | Uci | 2 | 5 |
| *Forest-fires* | Uci | 3 | 13 |
| *Wine* | Uci | 10 | 13 |
| *Yeast* | Uci | 3 | 8 |

Weka refere-se ao projeto Weka. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Clue é um pacote

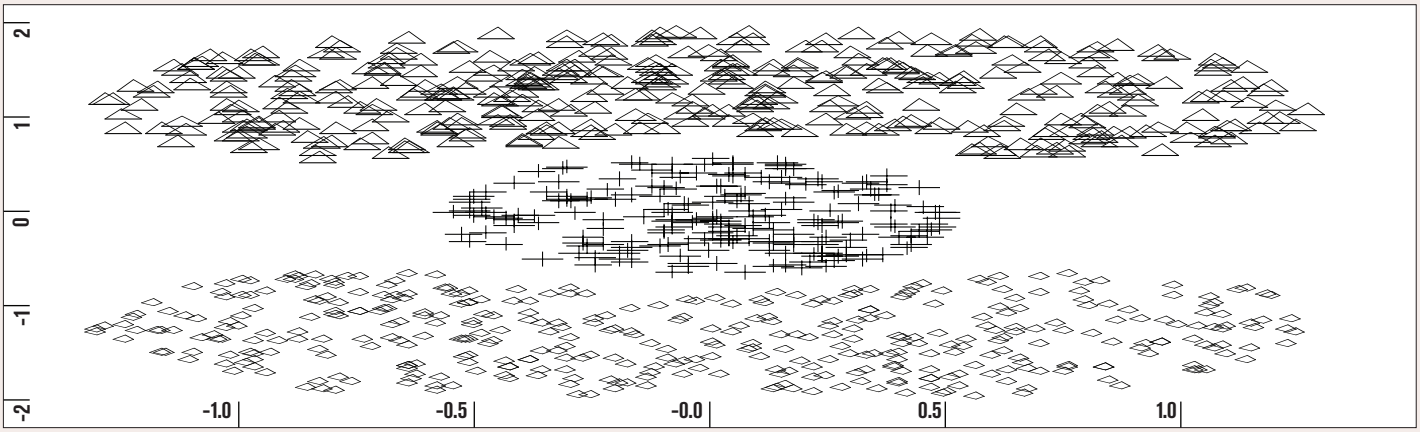


**FIGURA 4**
Dados de Cassini.

*(package)* do *software* estatístico R. Disponível em: <http://www.r-project.org/>. Qiu e Joe indicam que as bases de dados foram geradas com o método desses autores Qiu e Joe (2006a e b). Uci refere-se ao famoso UCI *Repository*. Disponível em: <http://archive.ics.uci.edu/ml/>.

Os dados de íris são, provavelmente, os dados mais famosos da estatística multivariada (é também a base mais acessada do UCI *Repository)*. Têm-se três grupos de íris (setosa, versicolor e virginica) com quatro variáveis, referentes às sépalas e pétalas (comprimento e largura).

A Figura 3 apresenta um diagrama de dispersão dos 150 casos da base, a partir das duas variáveis que mais discriminam entre os três tipos: os dados relativos às pétalas.

Nota-se que as íris setosas mostram-se em um grupo compacto e isolado, à esquerda e abaixo, no gráfico. Já entre as versicolor e virginicas, existe uma zona "cinzenta" que traz dificuldades ao algoritmo de agrupamento.

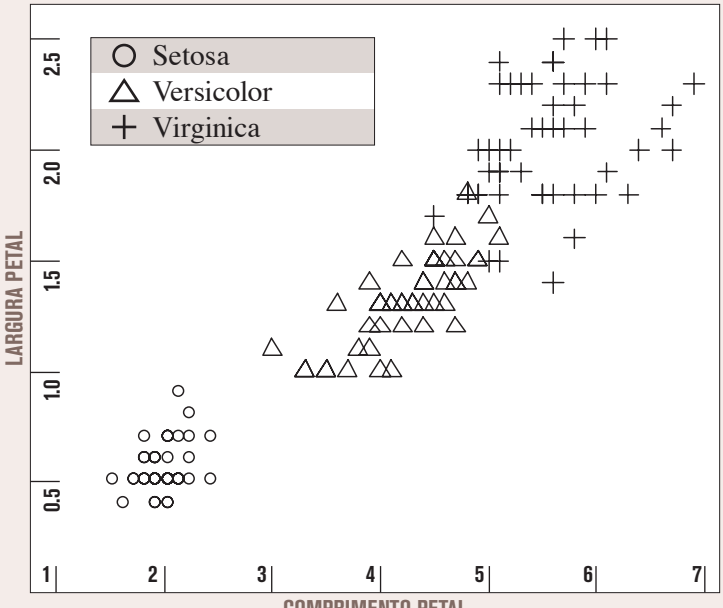


**FIGURA 3**
Dados de Íris.

Os dados de Cassini já apresentam outras dificuldades (Figura 4):

Aqui se têm três grupos, com duas variáveis, cujo formato alongado prejudica muito o desempenho do método K-médias, mas não o método K-medianas. Nesse conjunto de dados o método da Ligação simples se sai particularmente bem.

Dados 1 e Dados 2 têm uma estrutura mais compacta. São os dados mais assemelhados aos dados de estudos de segmentação dentre os catorze analisados. Um dos pontos que mais os diferenciam é que o segundo tem grupos mais separados, enquanto o primeiro tem menor grau de diferenciação (aproximando-se dos dados de um estudo de segmentação).

Os demais bancos de dados apresentam grande grau de dificuldade na *clusterização,* embora em alguns casos alguns métodos tenham apresentado um desempenho razoável.

## 4. AVALIANDO A QUALIDADE DAS SOLUÇÕES

Em todos os casos, é conhecido o grupo a que as observações pertencem, o que permite uma avaliação da qualidade de todas as soluções.

O índice utilizado para a avaliação da coerência entre a resposta correta e a solução sugerida pelos diversos métodos foi o Índice *Rand* ajustado de Hubert-Arabie, também chamado de *cRand*. Uma descrição detalhada desse índice, bem como uma comparação com diversos outros índices, pode ser obtida em Steinley (2004). Usualmente, um índice de 0,90 indica excelente qualidade de coesão; 0,80 um bom índice e 0,65 um valor moderado.

**TABELA 2**

Qualidade das soluções em cada método, medida pelo *cRand*.

| DATASET | FLX | KMN | DIV | SNG | WRD | FNY | KMD | SAN |
|---|---|---|---|---|---|---|---|---|
| Íris | 0,74 | 0,73 | 0,69 | 0,56 | 0,73 | 0,73 | 0,73 | 0,76 |
| Cassini | 0,08 | 0,53 | 0,51 | 1,00 | 1,00 | 0,97 | 0,93 | 0,91 |
| Dados1 | 0,84 | 0,95 | 0,69 | 0,00 | 1,00 | 0,96 | 0,93 | 0,86 |
| Dados2 | 0,82 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 |
| Cpu | 0,33 | 0,30 | 0,27 | 0,01 | 0,21 | 0,46 | 0,36 | 0,32 |
| *Segment-challenge* | 0,50 | 0,49 | 0,00 | 0,00 | 0,40 | 0,51 | 0,50 | 0,50 |
| *Breast-cancer* | -0,01 | 0,28 | 0,07 | -0,01 | 0,08 | 0,08 | 0,06 | 0,02 |
| *Segment-test* | (----) | 0,47 | 0,23 | 0,00 | 0,37 | 0,50 | 0,38 | 0,39 |
| *Ecoli* | (----) | 0,43 | 0,54 | 0,04 | 0,49 | 0,39 | 0,44 | 0,38 |
| *Cloud1* | 0,11 | 0,12 | 0,13 | 0,01 | 0,12 | 0,11 | 0,10 | 0,11 |
| *Mammographics* | (----) | 0,34 | 0,00 | 0,00 | 0,28 | 0,35 | 0,35 | 0,34 |
| *Forest-fires* | (----) | 0,00 | -0,03 | 0,00 | -0,04 | 0,00 | 0,00 | 0,00 |
| *Wine* | 0,81 | 0,66 | 0,51 | -0,01 | 0,58 | 0,60 | 0,54 | 0,65 |
| *Yeast* | (----) | 0,14 | 0,11 | 0,01 | 0,15 | 0,12 | 0,12 | 0,11 |

A Tabela 2 acima, indica a qualidade das soluções em cada método individualmente, onde os títulos das colunas seguem a mesma nomenclatura apresentada quando foram indicados, pela primeira vez, os métodos utilizados:

Nota-se que nas quatro primeiras bases de dados, o desempenho dos métodos é, em geral, bom. O que chama a atenção é que nem sempre um método é o melhor, o que ilustra o fato de que diferentes métodos exploram diferentes características da base de dados.

Como foi comentado anteriormente, houve casos em que o algoritmo de classes latentes (flx) não convergiu, o que é indicado na tabela com o símbolo de "(---)". No entanto, na base '*wine*' ele foi o melhor método, tendo também um excelente desempenho na base "íris."

O algoritmo de ligação simples (sng) foi muito bem em algumas bases e muito mal em outras. O algoritmo divisivo (div) sempre, ou quase sempre, foi o pior método. Chama a atenção a qualidade do desempenho dos métodos *Fuzzy* (fny), *Ward* (wrd), K-medianas (kmd) e *Simulated Annealing* (san).

Finalmente, a tabela 3 apresenta os "*Cluster Ensembles*."

As colunas representam, na ordem:

- O *Cluster Ensemble* com todos os métodos.
- O conjunto sem o método Divisivo.
- O conjunto sem o *single-linkage*.
- O conjunto sem o método Divisivo e o *single-linkage*.
- O conjunto sem o método Divisivo, o *single-linkage* e o de Classes latentes.
- O conjunto sem o método Divisivo e o de Classes latentes.

Nota-se a extrema estabilidade dos índices nos seis casos: a inclusão de um método com mau desempenho não degrada a solução.

Uma estratégia para a adoção desse método seria então, processar quaisquer dados com e sem o método de Classes latentes, não se aproveitando o conjunto total, mas apenas nos casos em que o algoritmo de classes latentes não convergir.

**TABELA 3**
Qualidade das soluções nos *"Cluster Ensembles"*.

| DATASET | CLE | CLE\DIV | CLE\SNG | CLE\DIV+SNG | CLE\DIV+SNG+FLX | CLE\DIV+FLX |
|---|---|---|---|---|---|---|
| Iris | 0,73 | 0,73 | 0,73 | 0,73 | 0,73 | 0,73 |
| Cassini | 0,83 | 1,00 | 1,00 | 0,97 | 0,97 | 1,00 |
| Dados1 | 0,90 | 0,91 | 0,91 | 0,91 | 0,90 | 0,90 |
| Dados2 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 |
| Cpu | 0,27 | 0,24 | 0,24 | 0,24 | 0,31 | 0,31 |
| *Segment-challenge* | 0,49 | 0,49 | 0,49 | 0,49 | 0,49 | 0,49 |
| *Breast-cancer* | 0,03 | 0,03 | 0,03 | 0,03 | 0,03 | 0,03 |
| *Segment-test* | (----) | (----) | (----) | (----) | 0,40 | 0,40 |
| *Ecoli* | (----) | (----) | (----) | (----) | 0,38 | 0,38 |
| *Cloud1* | 0,02 | 0,01 | 0,02 | 0,01 | 0,11 | 0,11 |
| *Mammographics* | (----) | (----) | (----) | (----) | 0,34 | 0,34 |
| *Forest-fires* | (----) | (----) | (----) | (----) | 0,00 | 0,00 |
| *Wine* | 0,53 | 0,53 | 0,54 | 0,53 | 0,51 | 0,51 |
| *Yeast* | (----) | (----) | (----) | (----) | 0,06 | 0,06 |

## 5. CONCLUSÃO

Os resultados apresentados indicam que a adoção dessa estratégia minimiza os riscos associados à escolha de um mau algoritmo para uma base qualquer. Os métodos não necessariamente "alavancam" a qualidade da solução, na medida em que nenhuma solução de *Cluster Ensemble* se mostrou melhor do que todas as soluções individuais.

Assim, o método funciona como uma espécie de "rede de segurança", já que um mau método, em geral, não prejudica a solução de conjunto para uma dada base.

Como forma de acelerar o processamento, uma estratégia seria utilizar os métodos de Classes latentes, K-medianas (ou *Simulated Annealing,* já que tiveram um desempenho semelhante), *Ward* e *Fuzzy*. O método Divisivo não superou os outros e o K-médias também não parece ter acrescentado muito.

Finalmente, fica também a observação de que uma pequena alteração pode ser avaliada, com o processamento de um método em várias versões: em cada uma delas se utilizaria uma seleção aleatória de um subconjunto das variáveis ou de componentes principais.

## 6. REFERÊNCIAS BIBLIOGRÁFICAS


BREIMAN, L. Random Forests. *Machine Learning,* 45(1), p.5-32, 2001.

HENDERSON, D.; JACOBSON, S. H.; JOHNSON, A.W. *The theory and practice of simulated annealing.* In: GLOVER, F., KOCHNBERGER, G. (Eds.) Handbook of Metaheuristics, Boston, Kluwer Academic Publishers, 2003.

HORNIK, K. A Clue for Cluster Ensembles, R. package. 2007. Disponível em: <http://www.cran.rproject.org/doc/vignettes/clue/clue.pdf>. Acesso em: 14 jun. 2008.

MINAEI-BIDOGLI, B.; TOPCHY, A.; PUNCH, W. *Ensembles of Partitions via Data Resampling, Proceedings da International Conference on Information Technology:* Coding and Computing (ITCC'04), v.2, p.188-191, 2004.

ORME, B.; JOHNSON, R. Improving *K-Means Cluster Analysis:* Ensemble Analysis Instead of Highest Reproducibility Replicates. White Paper. 2008. Disponível em: <www.sawtooth software.com>. Acesso em: 14 jun. 2008.

QIU, W.L.; JOE, H. Generation of Random Clusters with Specified Degree of Separation. *Journal of Classification,* 23(2), p.315-334, 2006a.

____. Separation Index and Partial Membership for Clustering. *Computational Statistics and Data Analysis,* 50, p.585-603, 2006b.

RETZER, J.; SHAN, M. *Cluster Ensemble Analysis and Graphical Depiction of Cluster Partitions,* Proceedings da 2007 Sawtooth Software Conference, 227f., p.239-250, 2007.

SOARES, L.; ESTEVES, W. *Consumer Drivers:* Estimando a Importância Derivada em Pesquisa de Mercado a partir de Random Forests. *3º Congresso Brasileiro de Pesquisa – Mercado, Opinião e Mídia,* ABEP. 2008.

STEINLEY, D. Properties of the Hubert-Arabie. Adjusted Rand Index, *Psycological Methods,* v.9, n.3, p.386-396, 2004.

UCI *Repository.* Disponível em: <http://archive.ics.uci.edu/ml/>. Acesso em: 14 jun. 2008.

WEDEL, M.; KAMAKURA, W. Market Segmentation – *Conceptual and Methodological Foundations,* Boston, Kluwer Academic Publishers, 2000.

WEKA projeto. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em 14 jun. 2008.

YEUNG, K.; RUZZO, W. Principal Component Analysis for Clustering Gene Expression Data, *Bioinformatics,* v.17 n.9, p.763-774, 2001.