Proceedings of the V Brazilian Conference on Neural Networks - V Congresso Brasileiro de Redes Neurais
pp. 379–384, April 2–5, 2001 - Rio de Janeiro - RJ - Brazil

"Ranking" de Bancos de Sangue Baseado em Mapas Auto-Organizáveis

Weber Martins[1,2], José Carlos Meira e Silva[1,3]
[1] UFG, Escola de Engenharia Elétrica, Grupo PIRENEUS, Praça Universitária, s/n,
[2] UCG, Mestrado em Psicologia, Área IV, Praça Universitária, s/n.
[3] Ministério da Saúde, DATASUS, SFCT-GO, Rua 82, 179, S. Sul,
Goiânia – Goiás, Brasil
weber@pireneus.eee.ufg.br, josec@fns.rgi.br

Abstract

There are applications that require ordered instances modeled by high dimensional vectors. Despite the reasonable quantity of papers on the areas of classification and clustering and its crescent importance, papers on ranking are rare. Usual solutions are not generic and demand specialist knowledge on the specification of the weight of each component and, therefore, the definition of a ranking function.

This paper proposes a generic procedure for ranking, based on unidimensional self-organizing maps (SOMs). Additionally, the similarity metric used by SOM is modified and automatically adjusted by a genetic search that seeks for the best ranking that matches the desired probability distribution provided by the specialist expectation. Promising results were achieved on the ranking of data from blood banks inspections.

1. Introdução

Apesar de pouco explorada, a ordenação envolvendo dados multidimensionais interessa a muitas áreas. Um exemplo é a necessidade, cada vez maior, dos "sites" de busca da Internet de implementar filtros cada vez mais poderosos, uma vez que a quantidade de informações na rede aumenta em ritmo acelerado. As simples contagens do número de ocorrências de palavras em páginas web já não selecionam e ordenam adequadamente, dado o grande universo de informações disponíveis. Soluções adequadas de ordenação na web deveriam usar dados multidimensionais (características do usuário, idioma, localização, históricos de buscas, etc). Outro exemplo de aplicação é usado, no presente trabalho, para avaliar o procedimento proposto. Trata-se da elaboração da ordenação ("ranking") dos bancos de sangue brasileiros, levando-se em conta a qualidade dos serviços realizados. A ordenação é resultado da ordenação de vetores de atributos levantados através de inspeções sanitárias efetuadas nos bancos de sangue.

Geralmente, um dado multidimensional modela algum fenômeno ou entidade da realidade, assumindo a forma de um vetor de atributos. Quais atributos do fenômeno serão escolhidos para representá-lo dependem da aplicação. Muitas vezes, os atributos escolhidos para a modelagem do fenômeno não são de nosso interesse direto. São apenas um meio de chegarmos aos atributos nos quais estamos realmente interessados. Um exemplo seria o dado multidimensional, usado em nosso trabalho, que modela uma inspeção sanitária em um banco de sangue. O interesse desta aplicação é avaliar os bancos de sangue de acordo com a qualidade dos serviços prestados. O atributo, "qualidade", porém, não é mensurável diretamente. Ele é obtido através da observação e análise de vários outros. Assim, o atributo no qual estamos realmente interessados é um atributo "latente". Neste sentido, como nos métodos da IRT (Item Response Theory) [10], a unidimensionalidade é requerida, ou seja, os atributos que modelam o fenômeno em estudo devem ser escolhidos de forma a contribuir para somente uma "variável" latente. Neste trabalho, a variável latente é a qualidade dos serviços prestados pelo banco de sangue. Os vetores de atributos dos bancos de sangue têm a dimensão igual a 77, isto é, para cada inspeção sanitária efetuada, existem 77 dados registrados que, de alguma forma, expressam o conhecimento da qualidade do banco de sangue.

Quando nos referimos a ordem (número ordinal), pensamos em uma seqüência que tem algum significado (ordem de classificação em um concurso, por exemplo). Assim, ao nos referirmos à colocação do Banco de Sangue de Quixeramobim no "ranking" dos bancos de sangue do Brasil, sabemos que existe algum critério (ou uma conjugação de critérios) que foi levado em conta para se chegar ao conhecimento da colocação ou posição no "ranking". Sempre que pensamos em ordem, pensamos em um critério unidimensional. A percepção humana de "ranking" é unidimensional. Se queremos fazer um "ranking" de instâncias modeladas por um dado multidimensional, devemos fazer uma redução de dimensionalidade aliada a algum tipo de regressão que, combinados, irão revelar o critério que será usado na formação do "ranking". A rede tipo SOM

(Self-Organising Map), conforme [6], apresenta estas duas características necessárias à formação de uma ordenação de instâncias representadas por dados multidimensionais.

O uso de redes neurais artificiais (RNAs) do tipo SOM na solução de problemas de ordenação de vetores de atributos de alta dimensionalidade não é comum. Problemas deste tipo são geralmente abordados usando-se classificadores probabilísticos e mesmo outras abordagens, conforme [4]. Kohonen, o criador do modelo, cita uma única referência a respeito do assunto em meio a aproximadamente 1700 referências [9].

Apesar do algoritmo SOM buscar a visualização de relacionamentos não lineares em dados multidimensionais, o desenvolvimento de uma "ordem" topológica (qualquer) nos mapas tipo SOM é um efeito já comprovado experimentalmente. Por outro lado, a prova analítica da ordenação resultante nestas redes não nos permite dizer que elas podem ser usadas na ordenação de instâncias representadas por vetores de alta dimensionalidade (a prova foi efetuada para o caso de redes unidimensionais e vetores de entrada de uma dimensão, valores escalares) [11]. O próprio Kohonen em [9] acha improvável a existência de estados ordenados convergentes em altas dimensões, ou seja, a existência de uma única ordenação quando os dados são multidimensionais. Em [1], através da observação de aspectos geométricos do fenômeno da ordenação descrito pelo teorema da ordenação de Kohonen [9], os autores demonstram que um único estado ordenado convergente em dimensões maiores que 1 tem probabilidade zero de existir, ou seja, é um evento impossível.

O fato de o mapa SOM não convergir voluntariamente para um (único) estado ordenado em espaços de entrada de altas dimensões requer a imposição de restrições externas para fazer uso das suas propriedades de redução de dimensionalidade e ordenação topológica. Estas restrições devem forçar a convergência ao estado onde a ordenação nos interessa e conduzir o processo de aprendizagem.

Outro aspecto a ser analisado diz respeito à métrica de similaridade usada no algoritmo SOM para se escolher a unidade vencedora. No algoritmo padrão, usa-se a distância euclidiana. A unidade que apresenta o vetor de pesos mais próximo (em termos da distância euclidiana) da entrada será o neurônio vencedor e terá o seu vetor de pesos atualizado (além de sua vizinhança). Em uma ordenação de instâncias, entretanto, nem sempre a distância euclidiana "pura" aponta para a ordenação desejada. Em um vetor de atributos, cada um deles pode ter um peso diferenciado no que se refere à ordenação. Por exemplo, o terceiro elemento de um vetor de atributos pode ser mais importante do que o primeiro elemento no contexto desejado da ordenação. A distância euclidiana pura não leva em conta esta ponderação de atributos. Assim, uma técnica mais genérica de ordenação deve considerar a ponderação dos elementos dos vetores de atributos que modelam as instâncias a serem colocadas em um "ranking". Neste trabalho, é sugerida uma ponderação dos atributos das amostras, usando-se um algoritmo genético para determinação automática dos valores usados na ponderação.

A função esperada de distribuição das amostras no "ranking" (fornecida pelo especialista no domínio) juntamente com as referências para a primeira e última instâncias da ordenação desejada (também fornecidas pelo especialista do domínio) são as restrições externas utilizadas para forçar a convergência da rede. A hipótese confirmada neste trabalho é que a função de distribuição e as referências fornecidas pelo especialista no domínio são suficientes para que a ordenação apresente convergência e para que a ordenação leve em conta o contexto desejado. Tal estratégia se justifica, tendo em vista que a determinação direta da importância relativa de cada atributo é processo controverso mesmo entre especialistas. Neste trabalho, usamos uma distribuição baseada em histograma obtido a partir da distribuição normal, dada à conhecida capacidade da mesma de modelar a realidade em diversas situações. Ao especialista, restou somente a determinação do melhor e pior casos.

2. O Algoritmo SOM

O algoritmo Self-Organizing Map (SOM) [6,7,8] é um método inspirado na biologia de certas regiões cerebrais (como a região de determinação das cores) para a construção de representações de dados a partir de um espaço de entrada de altas dimensões. Os dados são representados por protótipos implementados através de vetores de pesos. Tais vetores são associados aos neurônios (as unidades de processamento), de um espaço-imagem onde são definidas relações métricas entre os elementos. Para um conjunto de dados específico, o algoritmo SOM seleciona vetores de pesos e os atribui aos neurônios na rede. Os vetores de peso em função das coordenadas dos neurônios são chamados coletivamente de *mapa auto-organizável*.

3. Sistema Proposto

Este trabalho propõe o uso de RNAs tipo SOM unidimensionais para solução de problemas de ordenação de dados multidimensionais. Ele consiste em pesquisar e ajustar em uma população de métricas de distância, aquela que leve a rede SOM unidimensional a gerar um "ranking" no qual as diferenças de importância entre os atributos dos exemplos sejam levadas em conta. Isto é conseguido através da fixação dos pesos dos neurônios das extremidades da rede (com valores fornecidos pelo especialista no domínio) e do ajuste da distribuição das amostras no "ranking",

fazendo com que esta distribuição seja a mais próxima da esperada pelo especialista.

A pesquisa na população de distâncias é feita através de um algoritmo genético. Ele procura recombinar os melhores cromossomos (melhores distâncias) de acordo com o critério de qualidade estabelecido. No caso, este critério é a semelhança da distribuição do "ranking" gerado pela rede com a distribuição de referência arbitrada pelo especialista (neste caso, um histograma obtido a partir da distribuição normal).

O algoritmo do procedimento proposto é o seguinte:

1. Dimensionar a rede de acordo com o "ranking" desejado (neste caso, 10 neurônios/posições)
2. Gerar população inicial de distâncias (cromossomos) com valores randômicos escalares, na faixa informada pelo especialista
3. Enquanto a qualidade da distribuição não adequada faça:
   - 3.1 Fixar os pesos do último e do primeiro neurônio da rede através de amostras fornecidas pelo especialista no domínio. Os demais recebem valores randômicos escolhidos na mesma faixa de valores dos exemplos, conforme sugerido em [5]
   - 3.2 Recombinar distâncias da população (cromossomos)
   - 3.3 Treinar a rede com as amostras disponíveis, usando a distância (cromossomo) produto da recombinação
   - 3.4 calcular a qualidade do "ranking" gerado pela rede de acordo com a distribuição esperada
4. Treinar um número significativo de redes com a distância levantada e gerar "ranking" final através da média das colocações nas várias redes treinadas.

3.1. Métrica de Distância

A distância (ou medida de similaridade) usada no algoritmo SOM padrão é o quadrado da distância euclidiana entre a entrada e o vetor de pesos do neurônio. Sendo $\mathbf{E}$ o vetor de entrada, $\mathbf{w}$ o vetor de pesos e *n* a dimensão do vetor de entrada, a distância euclidiana "pura" é dada por:

$$d^2 = \Sigma (E_i - w_i)^2 \qquad (1)$$

onde:

E_i é o valor da entrada i; e

w_i é o valor do peso (conexão) i.

Neste trabalho, modificou-se esta distância para a seguinte forma conforme a equação (2) abaixo:

$$d^2 = \Sigma \beta_i * (E_i - w_i)^2 \qquad (2)$$

onde:

β_i é um fator de ponderação que indica a importância ("peso") do atributo i nos exemplos.

O algoritmo genético vai procurar o vetor β que impõe o melhor "ranking", de acordo com o critério da melhor distribuição dos exemplos (maior semelhança com uma distribuição arbitrada).

A fixação dos pesos dos neurônios das extremidades faz com que o mapa adquira um sentido de ordenação fixo, não permitindo inversões de sentido, comuns nos treinamentos das redes SOM unidimensionais. A fixação faz também com que a diferença de colocação (posição na ordenação) entre as duas referências (melhor e pior) indicadas pelo especialista seja naturalmente a maior do "ranking". A Figura 1 evidencia este efeito.

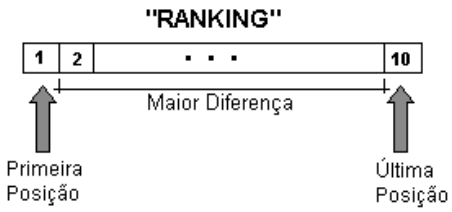

Figura 1: Efeito da fixação dos pesos dos neurônios das extremidades.

Obviamente, em uma situação onde são conhecidas as diferenças de importância relativas entre os atributos (onde o vetor β fosse conhecido a priori), o uso do algoritmo genético para ajustar estes parâmetros não é necessário. Geralmente, porém, estas ponderações são desconhecidas (e controversas). A técnica procura, portanto, ajustar os parâmetros, partindo do conhecimento da forma da distribuição final do "ranking". As diferenças de importância dos atributos são mapeadas através do ajuste da interpretação prática do espaço de entrada, ponderando a distância euclidiana. A ponderação funciona, pois, como um filtro. A Figura 2 mostra o esquema da técnica proposta.

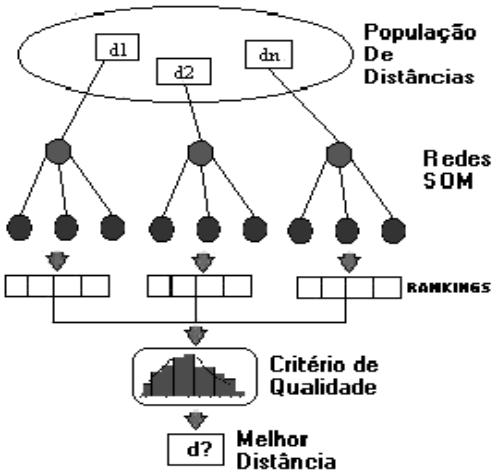

Figura 2: Esquema da técnica proposta.

3.2. Técnicas de Ordenação

Geralmente, os problemas de ordenação de dados multidimensionais são tratados através de ferramentas e métodos estatísticos. É possível, através de modelos de variáveis latentes e técnicas de análise de fatores, reduzir a dimensionalidade do espaço de entrada, chegando-se a algum critério unidimensional (*fator comum*). A partir daí, usando-se um método de regressão, poderia se chegar ao valor do critério latente para cada amostra e definir faixas de valores correspondentes às posições no "ranking". O problema é que, para se usar os modelos de variáveis latentes, os métodos de análise de fatores e as técnicas de regressão, muitas suposições a respeito dos dados e dos relacionamentos entre seus atributos devem ser feitas. Sabemos, no entanto, que os dados do mundo real nem sempre são "bem comportados". Eles não são sempre descritos por parâmetros estatísticos de baixa ordem, suas distribuições não são gaussianas e as relações funcionais entre seus elementos, freqüentemente, não são lineares. Além do mais, raramente, um dado de alta dimensionalidade (77 por exemplo) poderá ser reduzido a apenas uma dimensão (um fator na análise de fatores ou um componente na análise de componentes principais) sem grandes perdas de informação. Neste contexto, o procedimento proposto é mais genérico, porque não faz nenhuma suposição inerente sobre os dados e os relacionamentos dos seus componentes, exceto a unidimensionalidade no sentido de atender a busca de ordem. A restrição da forma da distribuição final do "ranking" não é fixa. Dependendo do "ranking" a ser feito, o especialista indicará a distribuição mais adequada para o problema, além das referências do melhor e pior caso.

4. Experimentos e Resultados

Os dados analisados para validação e estudo do procedimento foram levantados em inspeções sanitárias em bancos de sangue do Brasil. Ao todo, foram analisados dados de 85 inspeções. Cada inspeção foi registrada através de 77 itens. Cada item podendo receber os valores 1, -1 e zero, correspondendo, respectivamente, às resposta "sim", "não" e "não se aplica" ou "não informado". Os dados foram ordenados em três situações, cada uma gerando um "ranking". O primeiro "ranking" foi gerado usando um método estatístico (PCA -Principal Components Analysis e regressão) [2, 3]. O segundo foi gerado através de uma rede SOM sem o uso de ponderações e o terceiro usando a rede SOM e o ajuste da métrica de similaridade através do algoritmo genético.

4.1. Exp. 1:"Ranking" via Método Estatístico

Na obtenção do primeiro "ranking", foi usada a técnica de componentes principais para se chegar ao critério de ordenação. Tal técnica resultou em 23 componentes (maiores que um), que explicavam, juntos, 78,43% da variabilidade nos dados. O primeiro componente principal, que foi o escolhido para ser o critério de ordenação, representava 13,68% da variabilidade apenas. As faixas de valores usadas para delimitar as posições na ordenação e a forma do "ranking" (Figura 3) são mostrados na Tabela 1. Notar que a 3ª colocação recebeu 23 exemplos.

Tabela 1: Regra de Formação do Ranking

Posição no "Ranking"	Faixa de Valores
1ª	x > 1,5
2ª	1,0 < x ≤ 1,5
3ª	0,5 < x ≤ 1,0
4ª	0,0 < x ≤ 0,5
5ª	-0,5 < x ≤ 0
6ª	-1,0 < x ≤ -0,5
7ª	-1,5 < x ≤ -1,0
8ª	-2,0 < x ≤ -1,5
9ª	-2,5 < x ≤ -2,0
10ª	x ≤ -2,5

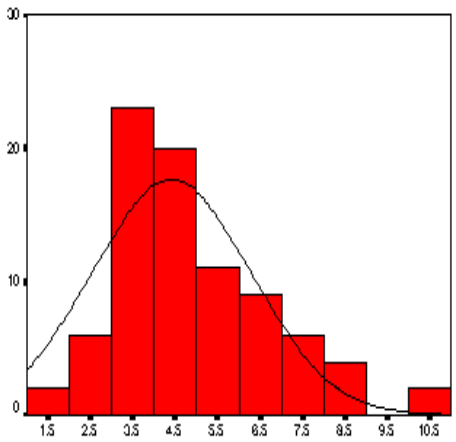

Figura 3: Histograma da distribuição dos exemplos no "ranking" usando método estatístico.

4.2. Exp. 2:"Ranking" via Rede SOM

Na obtenção deste "ranking" foi usada a rede SOM modificada de forma a fixar os pesos dos neurônios das extremidades com as referências dadas pelos especialistas. A ponderação e a conseqüente busca genética da melhor distribuição não foram usadas. Os parâmetros para o treinamento da rede na primeira fase foram taxa de aprendizado 0,9, raio de vizinhança 6 e número de passos 1044. Para a segunda fase foram taxa de aprendizado 0,02, raio de vizinhança 3 e número de passos 10440. A ordem de apresentação dos exemplos

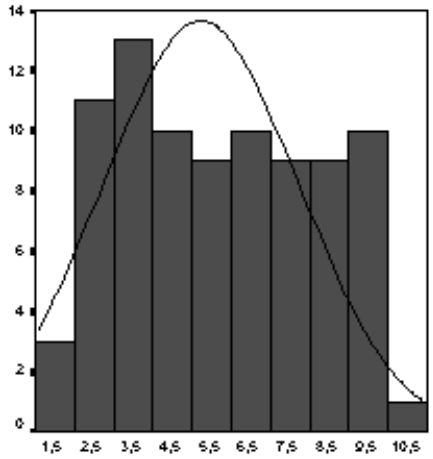

Figura 4: Distribuiçaon do "ranking" usando SOM padrão (distância euclidiana).

foi randômica e os valores dos parâmetros foram obtidos experimentalmente, levando-se em conta as sugestões em [9]. A forma obtida da distribuição das amostras é mostrada na Figura 4.

Verificamos nos experimentos, conforme já esperado com base nos estudos teóricos, que a cada treinamento a rede SOM unidimensional gera um "ranking" diferente (usando os mesmos parâmetros). Contudo, observamos também que a rede apresenta "convergência estatística" ou seja, à medida que aumentamos o número de redes treinadas, um determinado "ranking" tem uma probabilidade maior de acontecer. Com isso, para obtermos o "ranking" final precisamos treinar um número significativo de redes e, a partir delas, adotar as posições médias para cada exemplo. Neste trabalho o número de redes treinadas com esta finalidade foi de 1000.

4.3. Exp. 3:"Ranking" SOM - Ajuste Genético

Na obtenção deste "ranking" foi usada a rede SOM fixando-se os pesos dos neurônios das extremidades e alterando a métrica de similaridade conforme proposto neste trabalho. Os parâmetros de treinamento da rede foram os mesmos usados no experimento anterior. Os parâmetros do algoritmo genético foram os seguintes: tamanho da população inicial =100, taxa de mutação = 1%, "crossover" uniforme, inicialização dos genes dos cromossomos randômica, com valores na faixa 0 a 5. Da mesma forma que no experimento anterior, após obtidos os valores da ponderação através do algoritmo genético, foram treinadas 1000 redes e adotados os valores médios das posições dos exemplos. As formas da distribuição do "ranking", com valores iniciais dos cromossomos e após a convergência do algoritmo genético, são mostradas nas Figuras 5 e 6. Notar que a Figura 6 tem a forma mais próxima à normal.

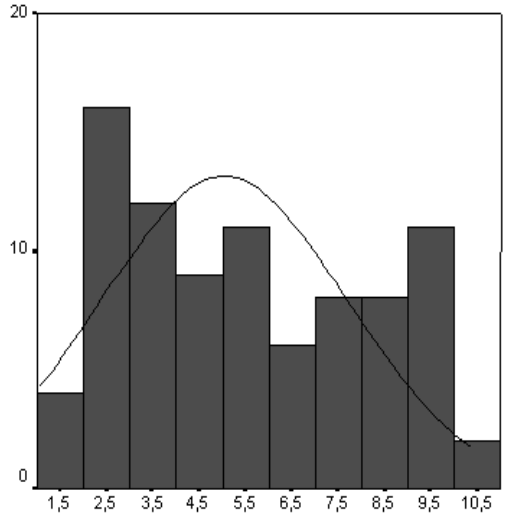

Figura 5: Histograma da distribuição do "ranking" usando SOM modificado, no início da busca genética. A 2ª colocação recebeu 17 exemplos.

4.4. Correlação entre os resultados

A base teórica sobre o assunto ainda não nos permite usar ferramentas especificamente desenvolvidas para comparação entre "rankings". Uma medida tradicional para tal fim é a correlação por postos de Spearman. Ela nos dá uma medida, pouca precisa, da semelhança entre dois "rankings". A Tabela 2 (ver abaixo) mostra a correlação de Spearman entre os resultados dos três experimentos.

Tabela 2: Quadro de Correlações

Experimento / Rankings	Correlação Spearman
1 x 2	0,937
1 x 3	0,860
2 x 3	0,916

Perceba que, pelos valores das correlações, a técnica SOM tradicional está a meio caminho das outras duas técnicas. A técnica estatística, devido à pouca expressividade do componente principal, não é recomendada nesta situação. A técnica SOM já apresenta um avanço, mas está baseada na distância euclidiana que assume a mesma importância para todo e cada atributo do vetor de características.

5. Discussão

Este artigo registrou a proposta de um sistema dedicado à descoberta de "rankings" para situações onde as técnicas da estatística tradicional não são adequadas. Adicionalmente ao uso dos mapas de Kohonen, desenvolveu-se, através do uso de algoritmos genéticos, uma técnica interessante para, ao mesmo tempo, atender às expectativas dos especialistas e

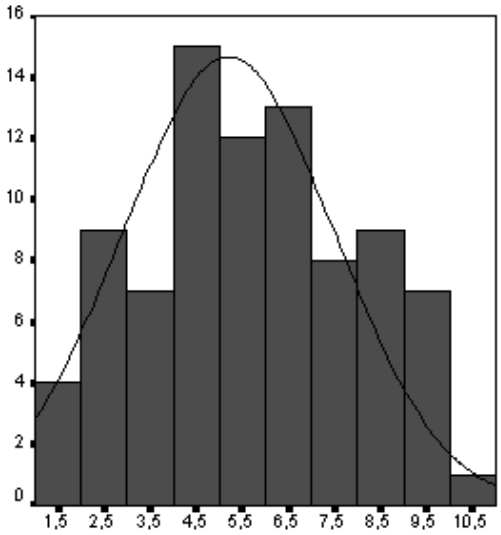

Figura 6: Histograma da distribuição do "ranking" usando SOM modificado, após convergência da busca genética.

encontrar o grau de importância (peso) de cada atributo do vetor de características que descreve o fenômeno.

A técnica apresentou resultados satisfatórios para o problema de ranqueamento dos bancos de sangue brasileiros pelo critério de qualidade, revelando a força de cada atributo contido no questionário de inspeção sanitária. A presente técnica abre também espaços no campo do uso do conhecimento de especialistas na condução de treinamento de redes neurais SOM, ou seja, trata-se de um método para o desenvolvimento de sistemas híbridos baseados em mapas auto-organizáveis.

Trabalhos futuros envolverão o uso da técnica proposta para estudo comparativo entre a redução de dimensionalidade resultante nas redes SOM e a conseguida através das técnicas estatísticas; bem como o estudo do uso da técnica como mecanismo de verificação de unidimensionalidade relacionada à Teoria da Resposta ao Item da Estatística tradicional.

Referências

[1] Budinich M., Taylor John G.(1995). On the Ordering Conditions for Self-Organizing Maps. Centre for Neural Networks - King's College. London.

[2] Carreira-Perpiñán M. A.(1999). One-to-Many Mappings, Continuity Constraints and Latent Variable Models. IEE Colloquium on Aplied Statistical Pattern Recognition. Birmingham-UK.

[3] Carreira-Perpiñán M. A.(1999). A Review of Dimension Reduction Techniques. Technical Report CS-96-09. Dept. of Computer Science – University of Sheffield.

[4] Cohen William W., Schapire Robert E., Singer Yoram (1999). Learning to Order Things. Journal of Artificial Intelligence Research 10 243-270.

[5] Fausset L.V.(1994). Fundamentals of neural networks. New Jersey. Prentice-Hall.

[6] Kohonen T. (1982). Analyses of a Simple Self-Organizing Process. Biol Cybern 44:135-140.

[7] Kohonen T. (1982). Self-Organized formation of topologically correct feature maps, *Biological Cybernetic*s, 43,pp59-69.

[8] Kohonen T. (1990). The Self-Organizing Map, *Proceedings of the IEE*E, 78, pp 1464:1480.

[9] Kohonen T. (2nd Ed. 1997). Self-Organizing Maps. Berlim. Springer-Verlag.

[10] Li H., Stout W.(1995). A Version of DimTest to Assess Latent Trait Unidimensionality for Mixed Polytomous and Dichotomous Item Response Data. Paper presented at the 1995 NCME Annual Meeting. San Francisco-CA.

[11] Erwin E, Obermayer, K. and Schulten K, (1992) Self-Organizing Maps : ordering, convergence properties, and energy functions, *Biological Cybernetic*s, 67, pp 47-55.