Busca eficiente em bancos de dados biométricos utilizando algoritmos de agrupamento sequencial

Jovani Antônio Maccarini, Luis Mariano Del Val Cura

Faculdade Campo Limpo Paulista (FACCAMP)
Campo Limpo Paulista – SP – Brasil

jovanimaccarini@gmail.com, delval@cc.faccamp.br

Abstract. *The method frequently used for templates searching in biometric databases is the expensive exhaustive searching. This paper proposes an alternative algorithm using the sequential clustering algorithm BSAS. The BSAS algorithm is used to organize the database in clusters considering the similarity measure among the data. The proposed algorithm reduces the searching space to the set of clusters most similar to the query template. This paper shows experimental results of the algorithm considering the variations of the size of searching space and parameters of the clustering algorithm BSAS.*

Resumo. *A busca de um descritor biométrico em um banco de dados frequentemente é realizada através do método de busca exaustiva que possui um alto custo computacional. Este artigo propõe um algoritmo eficiente para esta busca utilizando o algoritmo de agrupamento sequencial BSAS. O algoritmo de agrupamento é utilizado para organizar os descritores biométricos no banco de dados em grupos considerando a similaridade entre estes descritores. O algoritmo proposto reduz o espaço de busca a um subconjunto de grupos mais similares ao descritor de consulta. No trabalho são apresentados resultados experimentais que mostram o comportamento do algoritmo com variações no tamanho do espaço de busca e nos parâmetros do algoritmo de agrupamento BSAS.*

1. Introdução

O uso da biometria tem se expandido de forma dramática nos últimos anos como consequência do aumento da necessidade de identificação das pessoas, fundamentalmente em aplicações de segurança. Todos os dias são realizados milhares de cadastros dos descritores biométricos, e com este aumento do volume de informação, os problemas do armazenamento e a recuperação destes descritores tem se convertido em tarefas complexas.

Um descritor biométrico é representado como um vetor de características construído por um algoritmo biométrico a partir do processamento de um registro geralmente capturado na forma de imagem, como é o caso da face, impressão digital, íris, dentre outros. Uma característica do registro biométrico de um indivíduo é que ele será diferente cada vez que uma nova captura for realizada e, portanto, o descritor de cada registro será também diferente. Como consequência, para decidir se dois descritores biométricos pertencem ao mesmo indivíduo é necessário comparar estes

descritores através de uma função de similaridade. Esta função de similaridade pode ser utilizada em dois tipos diferentes de sistemas biométricos: sistemas de *verificação* e sistemas de *identificação* [Maltoni 2003]. Nos sistemas de *verificação*, um descritor de consulta se rotula com a suposta identidade de um indivíduo e é comparado com o descritor desse indivíduo na base de dados. Este sistema apresenta como saída a confirmação ou não da identidade. Em um sistema de *identificação*, um descritor de consulta é comparado com todos os descritores de indivíduos armazenados em uma base de dados. A saída deste sistema indica se o indivíduo existe ou não no banco de dados ou, então, pode ser o conjunto dos descritores de indivíduos (um ou mais) que mais se assemelham ao indivíduo no descritor de consulta. Em ambos os tipos de sistemas é utilizado um limiar de similaridade para tomar a decisão sobre a comparação dos descritores. Note-se que no caso de registros de baixa qualidade ou ainda por limitações do algoritmo biométrico, dois descritores de um mesmo indivíduo, quando comparados, podem gerar valores de similaridade baixos provocando erros de reconhecimento. Estes erros são próprios de qualquer sistema biométrico, isto é, espera-se em um sistema biométrico uma certa taxa de erros de reconhecimento.

Para implementar um sistema de *identificação* a solução mais simples é a busca exaustiva do banco de dados. Nesta busca, aplica-se a função de similaridade sobre o descritor de consulta e todos os descritores no banco de dados e seleciona-se aquele com maior valor de similaridade. Idealmente, o descritor selecionado deveria ser do mesmo indivíduo do descritor de consulta caso ele esteja representado no banco de dados.

Para realizar uma busca eficiente do descritor de maior similaridade, necessariamente o espaço de busca precisa ser reduzido. Métodos de acesso para buscas por similaridade eficientes em espaços multidimensionais têm sido amplamente pesquisados, mas é conhecida a sensibilidade desses métodos à alta dimensionalidade do espaço dos vetores como é o caso dos descritores biométricos. Uma outra alternativa para esta busca pode ser é o uso de algoritmos de agrupamento [Jain et. al. 1999]. Estes algoritmos organizam os dados em vários grupos ou classes, a partir de uma métrica de distância entre esses dados. Cada grupo deve possuir dados próximos segundo a métrica de distância e geralmente é caracterizado por um descritor, frequentemente calculado como o centróide dos dados do grupo. Para criar os grupos, os algoritmos de agrupamento podem precisar processar todos os dados repetidas vezes, isto é, processar os dados em vários passos.

Para utilizar um algoritmo de agrupamento para a busca eficiente em um sistema de identificação biométrico, gera-se inicialmente um conjunto de grupos com todos os descritores do banco de dados. Para este agrupamento é utilizada como métrica de distância alguma propriedade relacionada com os dados, por exemplo, a função similaridade biométrica. Para realizar a busca de um descritor de consulta o espaço de busca é reduzido a um subconjunto dos grupos com descritores mais similares a esse descritor de consulta. Sobre esse subconjunto de grupos realiza-se então uma busca exaustiva.

Algumas propostas na literatura exploram o uso de algoritmos de agrupamento para busca eficiente em bancos biométricos. Em [Iloanusi& Osuagwu 2011] e [Liu, Jiang & Kot 2007] o algoritmo de agrupamento K-means é utilizado para busca eficiente para descritores de impressões digitais. Uma aplicação para reconhecimento facial é apresentada em [Perronnin & Dugelay 2005] utilizando um agrupamento baseado em probabilidades para realizar busca eficiente. [Mehrotra et.al. 2009] por sua vez utiliza um algoritmo de agrupamento k-means fuzzy para aplicação em

reconhecimento biométrico da assinatura. Todas as propostas utilizam variações do algoritmo de agrupamento K-means [Jain et. al. 1999] que possui um alto custo de criação e não se adaptam facilmente à inclusão dinâmica de novos elementos.

Algoritmos de agrupamento sequenciais definem uma classe dos algoritmos de agrupamento que se caracterizam por realizar o agrupamento de forma rápida eficiente com uma ou poucas passadas pelo conjunto de dados [Theodoridis e Koutroumbas 2009]. Adicionalmente podem ser facilmente adaptados para a inclusão dinâmica de novos elementos. Estas propriedades refletem o comportamento desejado para um banco de dados: rápida construção da estrutura para busca eficiente e fácil inclusão de novos elementos no banco de dados.

Este artigo apresenta os resultados da pesquisa do uso do algoritmo sequencial BSAS (*Basic Sequential Algorithmic Scheme*) para a busca eficiente em bancos de dados de descritores faciais. Os experimentos realizados mostram o impacto na busca da redução do número de grupos explorados assim como com a variação dos parâmetros do método BSAS. O restante do artigo está organizado da seguinte forma. A seção 2 descreve o algoritmo de agrupamento sequencial BSAS, a seção 3 apresenta o algoritmo de busca eficiente utilizado nesta pesquisa e por fim a seção 4 descreve os experimentos realizados e os resultados obtidos.

2. Algoritmo BSAS - *Basic Sequential Algorithmic Scheme*

Algoritmos de agrupamento são representantes dos algoritmos de aprendizagem não supervisionado. Estes algoritmos classificam um conjunto de dados em classes a partir da identificação de similaridades compartilhadas pelos elementos de cada uma das classes. Para identificar estas classes estes algoritmos, em geral, realizam vários passos de processamento de todos os dados.

O algoritmo BSAS [Theodoridis & Koutroumbas 2009] pertence à classe dos algoritmos de agrupamento sequenciais e caracteriza-se por ser muito simples e eficiente na criação dos grupos. Ele recebe como entrada um conjunto de dados e realiza o agrupamento com um único passo de processamento. O comportamento deste algoritmo tem sido estudado com profundidade em [Real 2014] e [Real, Nicoletti & Oliveira 2013] mostrando taxas de precisão de classificação similares às alcançadas pelo algoritmo K-means, mesmo com um único passo de processamento.

O algoritmo recebe um conjunto de dados de entrada $\{x_1, x_2, ..., x_n\}$ e constrói o conjunto de grupos $\{C_1, C_2, ..., C_k\}$ tal que cada elemento x_i está em algum dos grupos. Para cada grupo C_i é definido um descritor T_i calculado como o centróide de todo o conjunto de elementos em C_i.

O algoritmo BSAS possui dois parâmetros que podem ser prefixados: um limiar de similaridade (Θ) que define a maior distância permitida entre os elementos de um mesmo grupo e; uma quantidade (q) que define o número máximo de grupos que podem ser criados. Como apresentado no Algoritmo 1, este algoritmo cria um primeiro grupo associado ao primeiro elemento x_1 do conjunto de dados e na sequencia vai adicionando os outros elementos utilizando os parâmetros Θ e q. Para adicionar um novo elemento x_i é determinado o grupo C_k mais próximo, isto é, com menor distância $d(x_i, T_k)$. Se esta distância é menor que o limiar Θ, o elemento é adicionado nesse grupo C_k. Caso contrário um novo grupo é criado e o elemento adicionado a esse grupo. Caso ocorra a

situação em que o valor de distância seja maior que Θ e, ainda, a quantidade de grupos ter atingido o valor *q* então este elemento será associado ao último grupo criado.

```
Algoritmo BSAS
Entrada:  Conjunto de dados x = { x1 , x2 , ...., xn },
          Parâmetros Θ, q
Saída: C conjunto de grupos C = { C1 , C2 , ....., Ck }

      m = 1
      C1 = {x1}
      Tk = {x1}
      Para i = 2..n
            Encontrar Ck tal que d(xi, Tk) = min1≤j≤md(xi, Tj).
            Se (d(xi, Tk)>Θ) e (m<q) então
      m = m + 1
      Cm = {xi}
      Tk = xi
            Senão
      Ck = Ck ∪ {xi}
      Recalcular Tk
```

Algoritmo 1. Algoritmo BSAS

Note-se que o algoritmo permite facilmente a incorporação posterior de novos descritores sem necessidade de processamento de todos os dados anteriores. Esta propriedade é importante porque um banco de dados biométrico terá necessariamente uma evolução dinâmica.

3. Algoritmo de busca eficiente

O algoritmo de agrupamento BSAS apresentado na sessão 2 é utilizado para organizar em grupos um banco de dados de descritores biométricos faciais, Como função de distância para o algoritmo BSAS é utilizada uma função de similaridade biométrica de forma tal que os descritores de imagens faciais similares devem ficar organizados nos mesmos grupos ou em grupos próximos. A idéia básica apresentada neste trabalho é explorar esta organização no banco de dados para realizar a busca de um descritor de consulta. Um algoritmo foi desenvolvido para reduzir a busca do descritor de consulta a um subconjunto dos grupos do banco de dados.

O algoritmo de busca eficiente utilizado neste trabalho recebe um descritor de consulta X_c, o conjunto de grupos C que corresponde ao banco de dados, e um parâmetro *p* chamado de taxa de penetração. A taxa de penetração define qual o percentual de grupos que deverão ser explorados na busca do descritor de consulta X_c. Quando a taxa de penetração é 100 o algoritmo se comporta como a busca exaustiva no banco de dados.

Como descrito no Algoritmo 2, o algoritmo procura primeiramente os *p* grupos mais similares a X_c. Esta similaridade é calculada utilizando a distância de X_{ca} de cada um dos centróides, que são os pontos de referência de cada grupo, isto é, são selecionados os *p* grupos com menor distância de X_{ca} de seus centróides. Uma vez determinados estes *p* grupos, os descritores em cada um deles são explorados de forma exaustiva para encontrar o descritor mais similar a X_c.

Algoritmo de busca eficiente
Entrada: Descritor de consulta x_s
Conjunto de grupos C = { C_1 , C_2 , ..., C_k } Banco de dados
Taxa de penetração p
Saída: x_m: Descritor mais similar a x_c em C

Encontrar o subconjunto S de grupos mais similares a x_s.
S = { C_{s1}, C_{s2}, ... , C_{sp} } tal que para qualquer grupo $C_k \notin S$,

$d(x_s, T_k) > d (x_s , T_{sk})$ para todo k = 1..p

Encontrar o descritor x_m em S mais similar a x_s

$d(x_s, x_m) = \min_{x_j \in C_{sk}} (d(x_s, x_j))$ para todo $C_{sk} \in S$

Algoritmo 2. Algoritmo de busca eficiente em grupos.

4. Experimentos e resultados

Para avaliar o algoritmo de busca proposto foram realizados um conjunto de experimentos utilizando 1.093 descritores da face obtidos pelo método EigenFaces (Turk& Pentland1991) em imagens do banco de dados FERET (Phillips, P. et. al. 2000). A função de similaridade utilizada sobre estes descritores foi à distância euclidiana.

Para a criação dos grupos no algoritmo BSAS, os descritores foram incluídos de forma aleatória. O impacto na busca de diferentes ordenações dos descritores durante a criação dos grupos não foi abordado neste trabalho.

Na busca por similaridade é aceitável o fato de que certa percentagem de buscas exaustivas não encontram, como descritor mais similar no banco de dados, um descritor do próprio indivíduo. Por esta razão, para os testes, foram utilizados 165 descritores de consulta pertencentes a indivíduos diferentes para os quais a busca exaustiva consegue encontrar descritores mais similares pertencentes aos mesmos indivíduos.

Para os experimentos foi medida a taxa de precisão A(p) quando modificada a taxa de penetração *p* do algoritmo de busca eficiente. A taxa de precisão A(p) mede qual o percentual dos descritores de consulta para os quais o algoritmo continua encontrando como mais similares descritores dos mesmos indivíduos.

Os resultados dos experimentos apresentados foram obtidos com a variação dos parâmetros Θ e *q* do algoritmo de agrupamento sequencial BSAS. A primeira busca para cada limiar foi pelo método de busca exaustiva com *p* = 100. Na sequência p foi sendo diminuído de 10 em 10 até que a última busca fosse realizada em apenas 10% do banco de dados.

Na Figura 1, resultado da busca com o primeiro processamento.

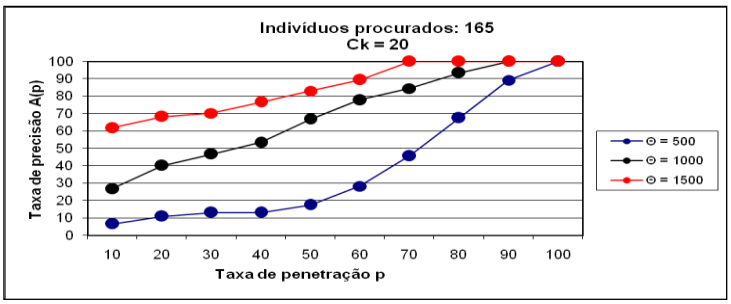

Figura 1. Resultado obtido na busca pelo agrupamento realizado com máximo de 20 grupos.

A Figura 1 apresenta os resultados obtidos pelo agrupamento realizado com máximo de 20 grupos (q) e valores de limiar de distância de Θ = 500, 1000 e 1500. Neste experimento o melhor resultado foi obtido com Θ = 1500, onde a precisão se mantém em 100% até uma taxa de penetração (p) de 70%.

Nos resultados apresentados na Figura 2, o agrupamento foi realizado com máximo de 50 grupos (q) e variação do limiar Θ = 500, 1000 e 1500. Neste caso o melhor resultado obtido manteve-se com o limiar Θ = 1500, e neste caso embora a precisão começa a cair com menor valor da taxa de penetração a queda é menos significativa para taxa de penetração menores.

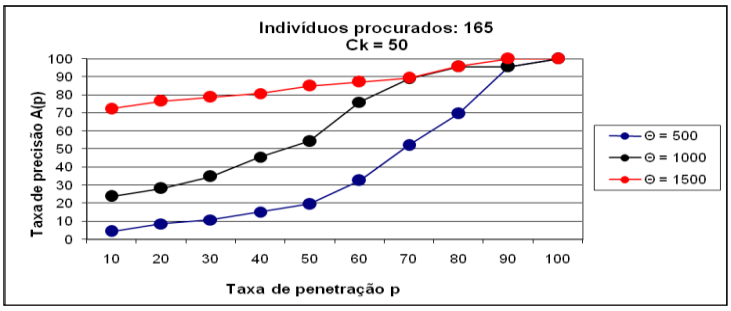

Figura 2. Resultado obtido na busca pelo agrupamento realizado com máximo de 50 grupos.

Finalmente, a Figura 3 mostra os resultados obtidos pelo agrupamento realizado com máximo de 75 grupos (q) e variação do limiar Θ = 500, 1000 e 1500. Neste caso com limiar Θ = 500 se garante resultados de precisão acima de 90% até uma taxa de penetração de 60% mas esta cai de forma acentuada quando a partir desse valor. No caso de limiar Θ = 1500 vemos que não garante a melhor precisão, mas garante a maior estabilidade desta precisão quando diminuída a taxa de precisão.

Os resultados sugerem que o parâmetro que mais influencia a precisão é o aumento do parâmetro Θ o que significa grupos menos coesos. Por outro lado, o aumento de q só gerou um resultado melhor unicamente com os valores maiores de Θ (1000, 1500). Precisam ser realizados novos testes aumentando o valor de q para verificar a partir de que valor a precisão começa a ser afetada. Note-se, no entanto, que um aumento de q significa também um maior custo computacional na fase de seleção dos grupos mais próximos.

Os resultados dos experimentos mostram um potencial de uso do algoritmo de busca eficiente quando uma aplicação deseja privilegiar a busca rápida aceitando certa perda de precisão. Em particular os resultados obtidos na Figura 3 com parâmetros q = 75 e Θ = 1000 e Θ = 1500 mostram comportamentos do algoritmo que se enquadram nessas situações.

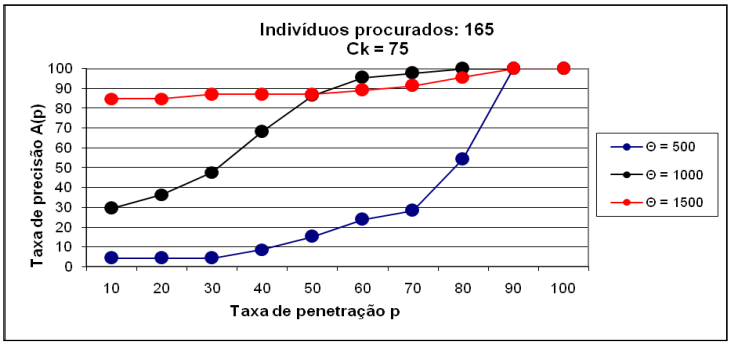

Figura 3. Resultado obtido na busca pelo agrupamento realizado com máximo de 75 grupos.

5. Conclusões

Esse artigo apresentou um algoritmo de busca eficiente em um banco de dados biométrico organizado em grupos gerados pelo algoritmo de agrupamento sequencial BSAS. A proposta considera a redução do espaço de busca, selecionando aqueles grupos mais similares ao descritor buscado. Forma apresentandos os resultados da pesquisa do desempenho do algoritmo para diferentes tamanhos do espaço de busca assim como para valores diferentes dos parâmetros do algoritmo BSAS utilizado. Os resultados dos experimentos mostram um potencial de uso do algoritmo de busca eficiente quando uma aplicação deseja privilegiar a busca rápida aceitando certa perda de precisão. Trabalhos futuros devem ser realizados refinando os parâmetros do algoritmo BSAS e devem ser também pesquisados outros algoritmos sequencias de agrupamento e outros conjuntos de dados biométricos gerados por métodos biométricos diferentes.

6. Referências

Iloanusi, O., Osuagwu, C. (2011). Clustering: Applied to Data Structurinh and Retrieval, International Journal of Advanced Computer Science and Applications, Vol 2, 11, 2011

Jain, A.K., Murty, M.N. & Flynn, P.J. (1999). "Data Clustering: A Review", ACM Computing Surveys, vol. 31.

Mehrotra, H., Kisku, D Radhika V., Majhi, B.; Gupta, P. (2009). Feature Level Clustering of Large Biometric Database IAPR Conference on Machine Vision Applications, May 20-22, 2009, Yokohama, Japan

Perronnin, F. & Dugelay, J.L. (2005). Clustering Face Images with pplication to Image Retrieval in Large Databases em Biometric Technology for Human Identification II, edited by Anil K. Jain,Nalini K. Ratha, Proc. of SPIE Vol. 5779

Phillips, P. et. al. (2000). The FERET evaluation methodology for face recognition algorithms, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(10), pp 1090-1104

Real, E. M., Nicoletti, M. C., Oliveira, O. L. (2013). The impact of refinement strategies on sequential clustering algorithms. Proceedings of the 2013 International Conference on Intelligent Systems Design and Applications (ISDA 2013). Piscataway, NJ, USA: IEEE Systems Man and Cybernetics Society, 2013. v. 1. pp. 47-52

Real, E. M. (2014). Investigação de algoritmos sequenciais de agrupamento com pré-processamento de dados em aprendizado de máquina. Dissertação de Mestrado

Theodoridis, S., Koutroumbas, K. (2009). Pattern Recognition, 4th ed., USA: Elsevier.

Turk, M. & Pentland. A. (1991). Face recognition using eigenfaces, Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–591.