Busca por redução de espaço em banco de dados biométricos utilizando algoritmos de agrupamento sequencial

Jovani Antonio Maccarini

Novembro / 2019

Dissertação de Mestrado em Ciência da Computação

Busca por redução de espaço em banco de dados biométricos utilizando algoritmos de agrupamento sequencial

Esse documento corresponde à dissertação de mestrado apresentada à Banca Examinadora para defesa da tese no curso de Mestrado em Ciência da Computação da Faculdade Campo Limpo Paulista.

Campo Limpo Paulista, 01 de novembro de 2019.

Jovani Antonio Maccarini

Prof. Dr. Luis Mariano Del Val Cura (Orientador)

Agradecimentos

Agradeço a minha esposa Édina, minha filha Betina, meus pais Lindomar e Zelinda, e demais familiares por todo o apoio recebido nessa caminhada.

Agradeço ao meu orientador Dr. Luis Mariano Del Val Cura pela dedicação em conduzir minha orientação.

Agradeço aos professores Dr. Osvaldo Luiz de Oliveira e Dr. Julio Cesar dos Reis que participaram da banca de defesa da minha dissertação de mestrado e a todos os professores do curso de PMCC, FACCAMP.

Agradeço a todos os meus amigos que fiz no decorrer do curso.

FICHA CATALOGRÁFICA

Ficha catalográfica elaborada pela
Biblioteca Central da Unifaccamp

M115b

Maccarini, Jovani Antônio

Busca por redução de espaço em banco de dados biométricos utilizando algoritmos de agrupamento sequencial / Jovani Antônio Maccarini. Campo Limpo Paulista, SP: Unifaccamp, 2019.

Orientador: Profº. Dr. Luis Mariano Del Val Cura Dissertação (Programa de Mestrado em Ciência da
Computação) – Centro Universitário Campo Limpo Paulista – Unifaccamp.

1. Algoritmos. 2. Busca por redução de espaço. 3. Dados biométricos. 4. Agrupamento por similaridades. 5. Agupamento sequencial. I. Del Val Cura, Luis Mariano. II. Campo Limpo Paulista. III. Título.

***Resumo**: Um descritor biométrico geralmente é representado por um vetor de alta dimensionalidade. Por esta razão, uma consulta biométrica frequentemente é realizada como a busca exaustiva no banco de dados biométrico do descritor armazenado mais similar ao descritor de consulta. A busca por redução de espaço pode ser incrementada organizando o banco de dado em grupos de descritores similares através de algoritmos de agrupamento. O algoritmo de consulta pode reduzir o espaço de busca ao subconjunto dos grupos mais similares ao descritor de consulta. Esta pesquisa propõe uma solução para esta busca utilizando dois algoritmos de agrupamento sequencial: Basic Sequential Algorithmic Scheme (BSAS) e Modified Basic Sequential Algorithmic Scheme (MBSAS). Estes algoritmos são conhecidos pelo baixo custo computacional para a criação e reorganização dos grupos. Neste trabalho são apresentados resultados experimentais com um banco de dados biométrico que mostram a eficiência e precisão da busca quando modificados o tamanho do espaço de busca e os parâmetros dos algoritmos de agrupamento.*

***Abstract**:* Biometric descriptors are usually represented as high-dimensional feature vectors. As a consequence, a query on biometric databases is often implemented as an exhaustive search, looking up the stored descriptor most similar to the query descriptor. In order to improve searching efficiency, databases can be organized by clusters of similar descriptors using clustering algorithms. Thus, the searching space could be reduced to the subset of the most similar clusters compared to the query descriptor. In this paper, we propose a database organization using two sequential clustering algorithms: Basic Sequential Algorithmic Scheme (BSAS) and Modified Basic Sequential Algorithmic Scheme (MBSAS). BSAS and MBSAS are low computational cost algorithms for clusters creation and reorganization. Experimental resultsshow searching efficiency and precision on both approaches using different searching space sizes and clustering parameters.

Sumário

Lista de Figuras

Lista de Algoritmos

Tabela

Abreviaturas e Siglas

BSAS - ***Basic Sequential Algorithmic Scheme***

MBSAS - ***Modified Basic Sequential Algorithmic Scheme***

FAR - ***False Aceept Rate***

FRR - ***False Reject Rate***

ABRE – Algoritmo de Busca por Redução de Espaço.

RMP – Registros médios percoridos

DV – Desvio padrão

Capítulo 1. Introdução

1.1. Contexto geral

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 de 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 (Pedrycz *et al*. 1997).

Um descritor biométrico (descritor – são as características biométricas extraídas apartir da capitura de uma imagem para reconehceimnto biometrico) 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 (Bonato *et al.* 2010). 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* (Jain et.al. 2011). 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 à 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 descritores 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 (Jain et.al. 2011).

Para implementar um sistema de *identificação* a solução mais simples é a busca exaustiva no 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.

Para dados complexos (características biométricas) é mais comum efetuar consultas por similaridade que consideram as características particulares de cada elemento. Estas consultas retornam os elementos do conjunto de dados que atendem a certos critérios de similaridade entre um ou mais elementos. Desta forma, o objetivo é encontrar elementos similares, pois, para dados complexos é extremamente difícil encontrar no conjunto de dados elementos distintos estritamente iguais (Hjaltason & Samet 2003).

Dentre as várias estruturas utilizadas em bancos de dados, métodos de acesso multidimensionais (MAMs) não atuam com boa performance na recuperação de dados fundamentalmente pela quantidade de dimensões. Estes métodos, também são conhecidos como estrutura de indexação espacial, organizam o espaço multidimensional de tal forma que somente algumas partes do espaço e um subconjunto dos objetos espaciais armazenados sejam considerados para contestar uma dada consulta espacial. Na medida em que o volume de dados cresce, aumenta o tempo e o custo do processamento de análise, exigindo mais recursos (Berry & Linoff 2004). Para aprimorar esta situação surgiu em 1994, a ideia de um processamento analítico online (*OnLine Analytical Processing* – OLAP) (Thomsen 2002).

Para (Pareek 2006), *OnLine Analytical Processing* é um tratamento para responder a consultas que são naturalmente multidimensionais de forma mais ágeis e que junta conceitos de *data warehousing* como ETL (*Extract, Transform and Load* – extração, transformação e carga de dados) além de relatórios relacionais e mineração de dados.

Em alguns tipos de dados multidimensionais, com uma quantidade relativamente pequena de dimensões, os métodos de ordenação conseguem resultados notáveis, No entanto, com dimensões grandes, a efetividade começa a ser um problema por causa do espaço multidimensional elevado.

1.2. Motivação

A cada dia mais e mais a biometria vem ganhando espaço considerável, tanto no reconhecimento de pessoas como na validação de acesso de pessoas em diferentes ambientes. Com este aumento surgem os problemas para recuperação de maneira rápida e com eficiência destas informações. Isso nos motivou a realizar esta pesquisa procurando pontos para minimizar o tempo de busca da biometria.

Uma alternativa para esta busca pode ser o uso de algoritmos de agrupamento (Jain *et al.*1999). Algoritmos de agrupamento 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.

Neste trabalho, 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.

Algoritmos de agrupamento sequenciais definem uma classe dos algoritmos de agrupamento que se caracterizam por realizar o agrupamento de forma rápida e eficiente com uma ou poucas passadas pelo conjunto de dados (Theodoridis & 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. Note-se que nestas técnicas a busca se resolve com algoritmos polinomiais e não logarítmicos.

O objetivo da pesquisa é avaliar as vantagens da busca em um banco de dados biométrico se este é organizado utilizando técnicas de agrupamento básicas. Nesta pesquisa serão avaliados dois algoritmos de agrupamento sequencial, BSAS (*Basic Sequential Algorithmic Scheme*) e MBSAS (*Modified Basic Sequential Algorithmic Scheme*), por serem métodos simples de interpretar e muito rápidos. O método de pesquisa dos dados no banco de dados aqui realizado, é atribuído pelo método de identificação, na qual não se sabe quem é o individuo que se está procurando. Essa pesquisa não entrará no detalhe de como é realizada a extração dos vetores de características de dados biométricos como, impressão digital, face, geometria da mão, íris, veias da retina, dentre outras.

Como parte do trabalho, foram implementadas as técnicas de busca e recuperação baseadas nestes algoritmos de agrupamento. Foi utilizado um algoritmo de extração de características baseado no método Eigenfaces aplicado sobre o banco de dados de imagens FERET. Sobre estas imagens foi calculada a taxa de penetração em várias etapas, indicando, em alguns resultados, boas taxas de busca sobre os elementos do conjunto de imagens. Como parte final definimos os resultados da pesquisa com relação a estes algoritmos.

1.3. Organização dos capítulos

O Capítulo 2 desta dissertação realiza uma visão geral dos algoritmos biométricos para a qual esta dissertação será apresentada. Na sequência, no Capítulo 3, realizamos uma visão geral dos algoritmos de agrupamento apresentando em que lugar aparecem os algortimos de agrupamento sequencial. O Capítulo 4 descreve, fundamentalmente, quais são os algoritmos que aplicamos nesta dissertação, algoritmos BSAS e MBSAS, com a descrição de cada um dos seus elementos. A seguir, no Capítulo 5, realizamos a fusão dos algoritmos de busca junto com os de agrupamento. No Capítulo 6 definimos os experimentos que vamos realizar junto com os valores resultantes. Finalmente, no Capítulo 7 apresentamos as conclusões.

Capítulo 2. Conceitos básicos sobre busca biométrica

A biometria é a forma de reconhecimento de um indivíduo por suas características: (a) fisiológicas (face, impressão digital, mão e olho) ou (b) Comportamentais (assinatura, voz e ritmo da escrita). Qualquer característica fisiológica ou comportamental pode ser utilizada para reconhecer um indivíduo (Jain *et al.* 1999),.

2.1. Características dos Sistemas Biométricos

Os sistemas biométricos têm por objetivo reconhecer um indivíduo com base em suas características físicas ou comportamentais e sua principal funcionalidade esta voltada para segurança, como por exemplo, bancos, setores públicos, pela policia, empresas, condomínios e etc.

2.2. Sistema Biométrico

Um sistema biométrico consiste de quatro módulos principais: captura, extração, armazenamento e comparação (Gonzalez 2000). Os módulos captura, extração e armazenamento fazem parte da etapa de cadastro, enquanto o módulo de comparação faz parte da etapa de reconhecimento.

2.2.1. Etapa de Cadastro

A Figura 1, mostra o fluxograma referente à etapa de cadastro do vetor das características no banco de dados envolvendo, captura, extração e armazenamento. Esta etapa é executada quando um novo indivíduo é incorporado no banco de dados ou seus descritores biométricos são atualizados.

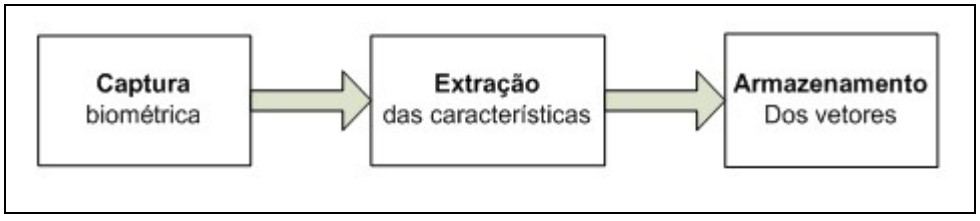

Figura 1. Fluxo de processamento em tempo de cadastro

2.2.1.1. Captura da amostra biométrica

A captura de uma amostra biométrica é o passo inicial do processo e pode ser realizada através de vários formatos, tais como, vídeo, foto, áudio e etc.

2.2.1.2. Extração das características da amostra capturada

A extração das características tem um papel importante, já que o processamento de comparação computacional de duas amostras é realizado através de uma representação numérica. Sendo assim a amostra capturada no inicio do processamento, passa pela extração de suas características ou conversão para um vetor de características (também conhecido como descritor) antes de qualquer tentativa de armazenamento ou comparação (Gonzalez 2000).

2.2.1.3. Armazenamento em banco de dados da amostra extraída

Depois de realizada a etapa de extração das características da amostra obtida na captura, esta amostra é armazenada no banco de dados para comparações futuras.

2.2.2 Etapa de Reconhecimento

Na Figura 2, a amostra capturada não é utilizada para cadastro no banco de dados como realizado na Figura 1, mas sim para reconhecimento do indivíduo. Este processo pode ser tanto para o método de *Verificação* quando *Identificação* (Ver Seção 2.3). O processo de comparação é responsável por comparar a amostra nova com as amostras cadastradas no banco de dados, apresentando no final da comparação se encontrou ou não o indivíduo apresentado ao sistema.

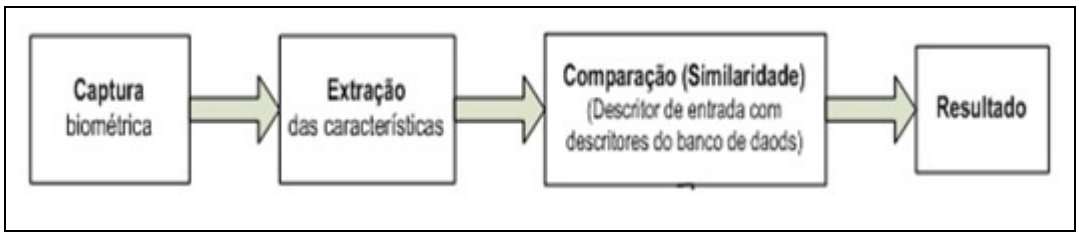

Figura 2. Fluxo de processamento em tempo de reconhecimento

2.2.2.1. Comparação

O processo de comparação é realizado no momento em que uma nova amostra é recepcionada. O processo de comparação recebe a amostra já processada na extração e, em seguida, compara-a com as amostras representadas por seus vetores de características cadastrados no banco de dados.

2.3. Diferença entre Verificação e Identificação

A etapa de reconhecimento biométrico pode ser realizada de duas formas: *Verificação* ou *Identificação, conforme* descrito nas Seções 2.3.1 e 2.3.2.

2.3.1. Reconhecimento por Verificação

A *verificação* é o método mais simples de reconhecimento biométrico e é comumente usada na autenticação de funcionários, em condomínios, por instituições financeiras, na autenticação de eleitores, escolas, etc. Nesta forma de reconhecimento o usuário informa ao sistema sua identidade e captura biométrica na qual o sistema vai realizar a verificação se esta identidade corresponde com o descritor armazenado no banco de dados. Um exemplo simples de autenticação por verificação pode ser a de um correntista de uma instituição financeira. No momento em que o indivíduo insere o seu cartão no leitor do caixa eletrônico é habilitado o sistema de leitura biométrica. Após a leitura e extração das características biométricas, o algoritmo associado compara o descritor da leitura com os descritores biométricos armazenados no banco de dados relacionado à sua agência e conta. No final da comparação o algoritmo atribui uma nota de similaridade entre as amostras e se a nota estiver dentro dos parâmetros que o sistema espera para autenticar, então o indivíduo é considerado como verdadeiro. Para medir a precisão deste processo em um algoritmo biométrico, são usadas duas taxas de ocorrências de erros, FAR (*False Aceept Rate*) e FRR (*False Reject Rate*). As taxas de ocorrência de tais erros são essenciais para se avaliar a eficiência do sistema. Tais taxas, definidas em (Jain *et al.* 2004) são:

- Taxa FAR, ou taxa de falsos positivos: probabilidade que um indivíduo não autorizado possa ser autenticado. Em caso de escala de semelhança, se a pessoa é

um impostor, mas sua pontuação correspondente é maior do que o limite, então ele é tratado como genuíno. Pode ser estimada como:

$$FAR = \frac{\text{número de falsas aceitações}}{\text{número de tentativas de impostores}}$$

• Taxa FRR, ou taxa de falsos negativos: probabilidade que um indivíduo autorizado seja incorretamente rejeitado, o sistema não consegue detectar uma correspondência entre o padrão de entrada e um modelo correspondente no banco de dados. Pode ser estimada como:

$$FRR = \frac{\text{número de falsas rejeições}}{\text{número de tentativas de genuínos}}$$

A Figura 3 apresenta duas distribuições de probabilidades de duas categorias (genuínos e impostores). Estas distribuições são constituídas a partir das notas que o algoritmo atribui a comparações genuínas e comparações impostoras. Elas descrevem a probabilidade de uma comparação ser genuína ou impostora, se escolhido o valor t como limiar. A alteração do valor atribuído ao limiar t, pode tornar o algoritmo mais seguro e reprovar um indivíduo genuíno ou permitir que um indivíduo impostor seja reconhecido como verdadeiro. Note que ao mover o limiar t para a direita, diminui-se a taxa de falsa aceitação (FAR), ao passo que a taxa de falsa rejeição (FRR) aumenta, e ao mover o limiar à esquerda, ocorre o contrário. O FAR e FRR dependem do valor atribuído ao limiar para tomar a decisão de aceitação ou rejeição (Jain *et al.* 2004).

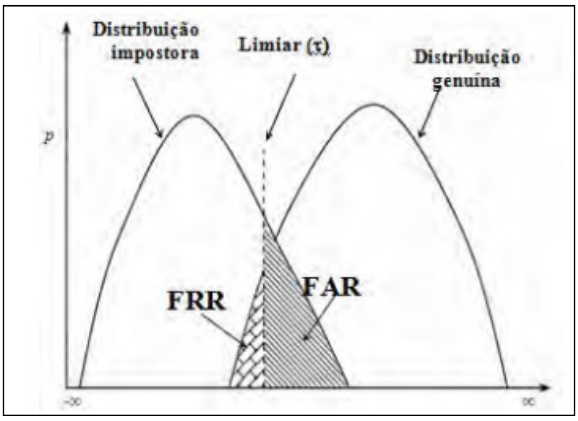

Figura 3. Distribuição das probabilidades de genuínos e impostores (Jain et al. 2004).

2.3.2. Reconhecimento por *Identificação*

No método de *identificação*, o método de busca mais simples é realizado pela busca exaustiva, onde é percorrido todo banco de dados comparando descritor de entrada com os descritores cadastrados no banco de dados. Portanto, o sistema realiza uma comparação para estabelecer a identidade de um indivíduo sem que o sujeito tenha fornecido qualquer tipo de identificação inicial. Embora os métodos tradicionais de reconhecimento pessoal, tais como senhas, chaves e crachás trabalham para o reconhecimento positivo, o reconhecimento negativo só pode ser estabelecido por meio de biometria (Jain *et al.* 2004). O reconhecimento positivo descrito significa que qualquer indivíduo pode se passar por verdadeiro mediante uma senha ou crachá de outro indivíduo. Para o reconhecimento biométrico esse indivíduo seria um impostor e o seu reconhecimento seria negativo.

Para o método de *identificação*, a taxa de erro FAR aumenta n vezes com relação à taxa de *verificação*. Na verificação como o processo é realizado de um-para-um a variação desta taxa é menor, variando com o grau de segurança estabelecido. Mas se tratando da identificação que realiza a comparação de um-para-muitos, essa taxa é n vezes maior, aumentando e muito a possibilidade de um falso indivíduo ser validado. (Jain *et al.* 2003).

Este método é muito utilizado no reconhecimento de pessoas procuradas pela justiça por meio da captação de imagens da pessoa usando de câmeras instaladas nas ruas, aeroportos, estações de metrôs e etc. Em um banco de dados estão cadastradas pessoas procuradas e essas câmeras espalhadas em pontos estratégicos, coletam amostras/descritores e comparam-nas com as amostras cadastradas no banco de dados.

2.4. Métricas biométricas, FAR, FRR, EER, FAR100 , FAR1000

Sistemas biométricos são projetados para validar o indivíduo confirmando que ele é realmente quem diz ser, ou rejeita-lo, considerando-o como impostor. Devido a variações na captura da amostra biométrica, tais como diferenças no posicionamento do traço biométrico, mudanças ambientais, deformações do traço biométrico na interação com o sensor, ruídos e má interação com o sensor, é praticamente impossível duas amostras de um mesmo indivíduo serem iguais. Por este motivo, adota-se uma nota que

quantifica o grau de similaridade existente entre o dado de entrada e o dado armazenado. Quanto maior for à nota para FRR, maior é a certeza que ambos referem-se à mesma identidade (Jain *et al.* 2003). Sendo assim, quanto maior o FAR, menor é a nota de FRR e maior será a chance de um indivíduo impostor ser aceito pelo sistema como verdadeiro.

O algoritmo dificilmente encontrará duas amostras idênticas, mesmo que ambas sejam de um mesmo indivíduo. Como a amostra sofre variações a cada captura, ainda que muito pequenas, os algoritmos trabalham com uma nota de comparação entre duas amostras para reconhecer ou não um indivíduo. A partir de uma nota determinada pelo algoritmo comparando a amostra recebida com as amostras do banco de dados, informando se aquele indivíduo é genuíno. A situação em que as duas taxas FAR e FRR têm a mesma nota, é chamado de *Equal Error Rate* (EER), o ponto da curva em que a taxa de FAR e FRR se igualam. Quanto menor for o EER mais preciso é o algoritmo (Ribeiro *et al.* 2010). Também entram os parâmetros FAR100 e FAR1000.

O FAR100 é o valor esperado para a taxa de FRR quando a taxa de FAR é igual a 1/100. Esta medida é útil para caracterizar a assertividade dos sistemas de reconhecimento biométrico. O FAR1000, é similar à medida de FAR100, onde o valor esperado de FRR, quando a taxa de FAR, é igual a 1/1000. Desta maneira o sistema é ainda mais rigoroso (Cappelli *et al.* 2006).

A taxa de erro FMR descreve a probabilidade de um indivíduo não ser identificado em um banco de dados por causa de que comparações impostoras produzem valores de similaridade maiores que as comparações genuínas. Por esta razão, é comum em sistemas reais que o reconhecimento por *identificação* devolva os N indivíduos mais simulares com o descritor de consulta e o usuário decide se entre eles se encontra o indivíduo procurado.

2.5 Métricas de similaridade

A métrica de similaridade é utilizada no meio computacional para encontrar objetos do mesmo domínio através da avaliação da similaridade, através de funções que medem a distância entre um ponto e outro.

Estas funções caracterizam-se a algoritmos computacionais que recebem dois objetos de um mesmo tipo e devolvem o valor da distância entre eles. Geralmente, uma função de distância atende d(x, y) = d(y, x) de forma não negativa. Além destas, se a desigualdade triangular: d(x, y) <> d(x, z) + d(y, z) também for respeitada, então, a função é considerada uma função de distância métrica, ou simplesmente métrica (Sousa *et al.* 2001).

O tempo de comparação entre dois descritores é muito alto dependendo da métrica que foi utilizada e, no caso dos objetos multi-atributos, da quantidade de atributos que os compõem. Logo, a consulta por similaridade em bancos de dados complexos e grandes pode ser considerado o tempo total para comparar todos os objetos. Outras técnicas para agilizar as consultas por similaridade, são baseadas na redução do número de atributos a serem indexados, como por exemplo, redução de atributos por mapeamento e transformação espacial e seleção de atributos relevantes (Sousa et al. 2002).

2.6. Taxa de penetração e taxa de precisão

Quando aplicamos técnicas de busca eficientes de redução do espaço de busca, a taxa de penetração p define qual o percentual do banco de dados que será explorado nessa busca. Quando a taxa de penetração é 100% o algoritmo se comporta como a busca exaustiva no banco de dados, comparando o descritor de entrada com os descritores armazenados. O algoritmo procura primeiramente os p grupos mais similares a Xc (descitror referência ou centroide). Esta similaridade é calculada utilizando a distância de Xca 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 Xca 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 Xc. A métrica utilizada para verificar a similaridade entre os descritores é pela distância euclidiana (formula utilizada tanto para o agrupamento quanto a busca).

A taxa de penetração, vai nos permitir visualizar o ganho ou perca na redução do espaço de busca, retornando a taxa de precisão A(p) (Capítulo 5). É claro que, em alguns pontos a redução deste espaço, resultados ruins serão encontrados e em outros percentuais de penetração, resultados bons ou aceitáveis serão encontrados, dentro da

margem de erro. Com isso, ganha-se em custo de processamento. Quando se fala em bando de dados biométricos com reconhecimento por identificação, imaginam-se grandes bancos biométricos e com consultas frequentes.

Capítulo 3. Conceitos de classificação e algoritmos de agrupamento

Neste capítulo serão apresentados os conceitos básicos de classificação supervisionada e não supervisionada, Tipos de algoritmos agrupamento de dados biométricos em grupos, algoritmos de agrupamento e revisão de trabalhos na literatura.

3.1. Classificação Supervisionada e não Supervisionada

3.1.1. Supervisionada

A classificação supervisionada consiste na identificação prévia das classes de informação, chamadas áreas de treinamento, que nada mais são do que representações do comportamento médio das classes que serão mapeadas automaticamente (Novo 1992).

Esse tipo de classificação é utilizado quando se possui algum conhecimento sobre as classes que devem ser representadas pelo computador, chamados de dados rotulados, podendo ser classificados por ordem crescente, decrescente numérica ou alfanumérica.

3.1.2. Não supervisionada

O método de classificação não supervisionado é usado para construção de grupos não rotulados. Por definição não se sabe a qual classe um objeto pertence e então são utilizados métodos de agrupamento.

Estes métodos de agrupamentos são baseados na similaridade entre um conjunto de dados, de tal maneira que os grupos obtidos sejam os mais homogêneos possíveis.

Há uma dificuldade em agrupar dados sem um padrão pré-definido, ou seja, dados não rotulados. A identificação de grupos similares de amostras a partir de um conjunto de dados é um passo crucial em muitas aplicações de análise de dados. As amostras são geralmente representadas por vetores de características, cuja semelhança entre eles depende de uma função de distância.

Uma das técnicas de agrupamento não supervisionado é a opção de grupos, através de métodos métricos identifica se os descritores mais similares. Quando

identificado dois pontos similares entre si, este então é associado ao grupo mais similar e as classes tendem a cair um ponto médio entre eles (Eastman 2006), de tal maneira que os grupos obtidos sejam os mais homogêneos possíveis.

3.2. Tipos de algoritmos de agrupamentos

Os algoritmos de agrupamento são divididos em categorias, cada algoritmo possui características próprias e geram resultados de agrupamento diferente entre eles. Abaixo serão apresentadas as diversas categorias dos algoritmos de agrupamento (Theodoridis & Koutroumbas 2009):

Algoritmos sequenciais. Estes algoritmos geram um único agrupamento, ou seja, os dados são apresentados ao algoritmo de uma única vez, grupando os vetores em grupos. Este tipo de algoritmo é considerado simples e rápido. O resultado final dependente da ordem em que os vetores são apresentados ao algoritmo.

Agrupamento hierárquico. Algoritmos de aglomeração e algoritmos divisíveis:

Algoritmos de aglomeração. Estes algoritmos geram uma sequência de grupos e na etapa seguinte, fundindo dois grupos em um. Sempre que uma nova etapa é realizada e o processo se repete, fundindo os grupos anteriores.

Algoritmo divisível: O algoritmo divisível é o oposto do algoritmo de aglomeração, ou seja, a cada processamento ele divide um único grupo em dois.

Algoritmos de agrupamento com base na otimização da função custo. Essa categoria de algoritmo é representada por uma função de custo, j, em termos de, m, agrupamentos. Geralmente, o número de grupos é mantido. A maioria desses tipos de algoritmos usam cálculos diferenciados e terminam quando um local ótimo de, j, é determinado.

Algoritmos de agrupamento encadernados: Estes algoritmos realizam agrupamentos sem que seja passado um número de grupos a ser criados.

Algoritmos de agrupamento genéticos: Utilizam uma população inicial de possíveis grupos e geram novos grupos, que, em geral, contêm melhores agrupamentos do que os das gerações anteriores, de acordo com um critério pré-especificado.

Algoritmos baseados em densidade: Estes algoritmos visualizam os aglomerados como regiões do espaço dimensional que são "denso". Deste ponto de vista, há uma afinidade com os algoritmos de busca.

Stochastic relaxation methods: Este é um métodos que garantam, sob certas condições, a convergência em probabilidade para o agrupamento integralmente ótima, respeito os critérios pré-especificados.

Algoritmos de agrupamento Valley-seeking: São algoritmos que tratam os vetores de características como instâncias de uma variável aleatória.

Subspace clustering algorithms: Estes algoritmos são adequados para o processamento de conjuntos de dados de alta dimensionalidade. Em algumas aplicações, a dimensão do espaço de características pode ser da mesma ordem de alguns milhares.

Algoritmos de aprendizado competitivo: Estes algoritmos não empregam funções de custo. Eles produzem vários agrupamentos sempre para o mais similar, de acordo com uma distância métrica. Representantes típicos desta categoria são algoritmos de aprendizado competitivo básico e o algoritmo de aprendizado *leaky*.

Os algoritmos baseados em técnicas de transformação morfológica: Algoritmos que utilizam transformações morfológicas, a fim de conseguir uma melhor separação dos aglomerados envolvidos.

3.3. Agrupamento de dados biométricos em grupos

Uma enorme quantidade de informações biométricas é inserida todos os dias em bancos de dados. Com essa enorme volume de dados a consulta tende a ficar mais lenta, principalmente ao se tratar de dados não supervisionados. Os descritores de características biométricas com processamento de busca pelo método de *identificação* não possuir um índice ligado ao registro e faz com que os dados armazenados precisem de alguma maneira serem agrupados a fim de facilitar a consulta posteriormente.

Agrupamento de dados é uma técnica essencial utilizada na estrutura subjacente em um conjunto de dados não supervisionados, agrupados por padrões de similaridade. Esse processo de agrupamento acelera a recuperação de uma informação no banco de

dados, onde, a distância é extraída como um recurso de busca auxiliar, tornando assim a busca mais eficiente e mais rápida (Manhua Liu *et al.* 2006).

3.4. Algoritmos de agrupamento

Uma solução para o problema de agrupar dados não supervisionados é o uso de algoritmo de agrupamento sequencial, organizando-os em grupos. Grupo é um conjunto de dados agrupados também chamados de grupos. Cada algoritmo de agrupamento tem suas particularidades na forma de agrupar ou de organizar os dados. Lembrando que os dados aqui tratados são vetores de características, na qual não é possível realizar o agrupamento na forma tradicional, ascendente ou descendente (dados não supervisionados).

A pratica comum entre algoritmos sequenciais é a criação de grupos. Ela representa uma das principais etapas de processos de análise de dados, denominada análise de grupos (Jain *et al.* 1999). Na Figura 4, uma representação de agrupamento de vetores em grupos no formato métrico por menor distancia entre o ponto central e os pontos ao seu arredor.

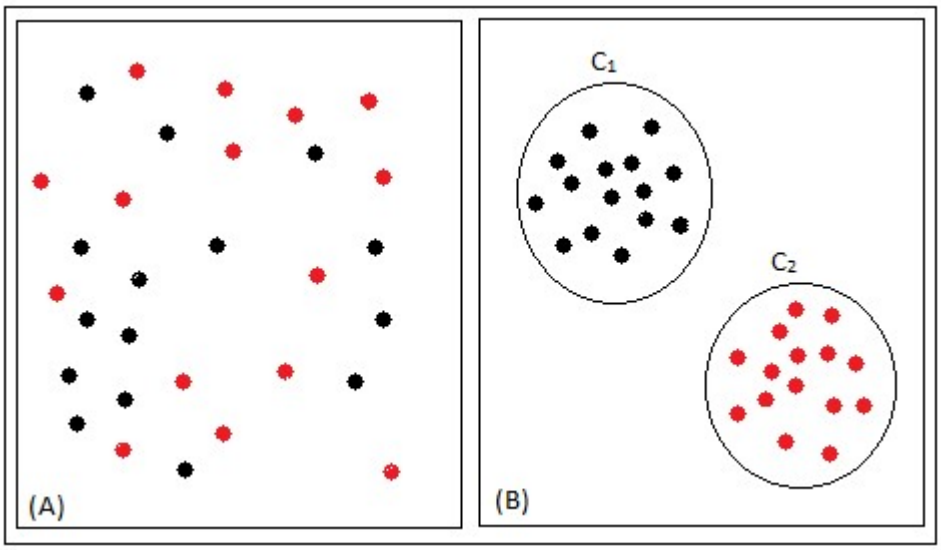

Figura 4. Representação de agrupamento.

Observando a Figura 4-A representa os vetores de características sem nenhum agrupamento, na 4-B, a representação após passar pelo processo de agrupamento por similaridade, onde foram criados dois grupos, C_1 e C_2 unindo cada descritor no grupo mais representativo. O que vai definir onde ele será agrupado é a métrica entre o vetor

lido e o vetor central dos grupos. Se a distância encontrada entre o descritor e o C_1 for menor que a distância entre o descritor e o C_2, então, este descritor será incluído no C_1. Pode ocorrer casos em que a distância encontrada entre eles seja a mesma, neste caso, cada algoritmo de agrupamento segue um padrão próprio pré definido no algoritmo.

Para os algoritmos aqui estudados, é fornecido pelo usuário o parâmetro referente à quantidade máxima de grupos que o algoritmo poderá criar para organizar os descritores no banco de dados. O grupo criado possui um centróide (cada grupo possui um ponto central referencial, conhecido na literatura como centroide, este centroide é definido toda vez que um descritor novo tem sua métrica acima do(s) centroide(s) já criado(s) anteriormente. No inicio do processamento o primeiro descritor lido torna-se o primeiro centroide) e para cada centróide um grupo de dados ou descritores são agrupados a este centroide. Estes grupos são formados pela similaridade do descritor de entrada com o centróide. Por exemplo, no inicio do processamento o usuário informou que pretende criar 10 grupos, cada grupo possui um centróide. A base a ser organizada possui 1000 descritores, neste caso estes 1000 descritores serão aglomerados no máximo em 10 grupos. Porém, não significa que cada grupo receberá 100 descritores cada, tudo vai depender da similaridade entre as eles, alguns grupos possivelmente receberão mais descritores que outros. O resultado final depende do algoritmo específico e os critérios utilizados. Assim, um algoritmo de agrupamento é um processo de aprendizagem que tenta identificar as características e atribuir ao grupo mais similar a ele. Um dos problemas consiste que, em alguns casos um descritor, x, deveria ser agrupado ao grupo C_n, porém, por algum motivo, o descritor x, foi agrupado ao grupo C_m. Em situações como essa, a possibilidade do descritor x, não ser encontrado em tempo de pesquisa é grande, pois, algumas técnicas de pesquisa não percorrem o banco de dados por completo, devido a sua volume ser elevada e o tempo de resposta ser muito lento, ficando inviável manter o sistema.

3.5. Trabalhos na literatura

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. O K-means é um algoritmo simples e eficiente e

tem sido adaptado e usado em muitos domínios de problemas ((Das 2003), (Maurya et al. 2011), (Tatiraju & Mehta 2008) e (Guan et al. 2013)). Embora ele possa terminar o processamento e não necessariamente encontra melhor a configuração para os grupos além de ser também bastante sensível ao conjunto de centróides inicialmente escolhidos (Bottou & Bengio 1995).

Uma aplicação para reconhecimento facial é apresentada em (Perronnin & Dugelay 2005) utilizando um agrupamento baseado em probabilidades para realizar a busca eficiente.

Mehrota (Mehrotra *et al.* 2009) por sua vez utiliza um algoritmo de agrupamento K-means e fuzzy para aplicação em reconhecimento biométrico da assinatura. Todas as propostas utilizam variações do algoritmo de agrupamento K-means que possui um alto custo de criação e não se adaptam facilmente à inclusão dinâmica de novos elementos (Jain *et al.*1999).

Os autores (Perronnin & Dugelay, 2005), realizaram uma avaliação sobre a eficácia de agrupamento para recuperação rápida de descritores de imagem facial. A abordagem estudada baseia-se no agrupamento dos descritores faciais em grupos. Utilizadas duas técnicas baseadas na classificação para reduzir o número de comparações ao pesquisar um banco de dados. A primeira abordagem faz uso de duas medidas complementares para a distância entre os descritores:

- A primeira distância tem uma baixa precisão, mas requer pouca computação, é executada em todo o conjunto de N-melhores candidatos e são mantidos.
- A segunda distância tem uma alta precisão, mas requer mais computação, por exemplo, ao problema de autenticação biométrica multimodal de indivíduos (Hong & Jain, 1998).

A segunda abordagem consiste em particionar o espaço agrupando o conjunto de dados. Quando um novo descritor é adicionado ao banco de dados, é calculada a distância entre o descritor de entrada e todos os descritores associados ao centróide do grupo. Quando um descritor de consulta é pesquisado, a primeira etapa consiste em encontrar o grupo mais próximo e o segundo passo envolve a computação das distâncias

entre os descritores de consulta e os descritores agrupados ao grupo correspondente (Jain & Pankanti, 2001).

Capítulo 4. Algoritmos de agrupamento sequenciais

Este capítulo trata sobre agrupamento biométrico em grupos, algoritmos de agrupamento, tipos de algoritmos de grupamento, algoritmos sequenciais BSAS (Basic Sequential Algorithmic Scheme) e MBSAS (Modified Basic Sequential Algorithmic Scheme).

4.1. BSAS - Basic Sequential Algorithmic Scheme

Algoritmos de agrupamento são representantes dos algoritmos de aprendizagem

não supervisionados. 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 {x1, x2, ..., xn} e constrói o conjunto de grupos {C1, C2, ..., Ck} tal que cada elemento xi está em algum dos grupos. Para cada grupo Ci é definido um descritor Ti calculado como o centróide de todo o conjunto de elementos em Ci.

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 agrupamentos que

podem ser criados. Durante o processamento, o primeiro descritor do conjunto de dados de entrada x1 torna-se o centróide T1, e recebe uma marca como sendo o ponto de referencia do grupo C1. No processamento do próximo descritor x2, se este atender ao parâmetro de similaridade passado ao algoritmo BSAS, x2 será inserido ao grupo C1 como um descritor comum. Caso contrário, x2 tornar-se o centroide T2 do grupo C2 e assim sucessivamente com os demais descritores contidos no conjunto de entrada.

Como apresentado no Algoritmo 1, linhas (1)-(3), este algoritmo cria um primeiro agrupamento associado ao primeiro elemento x1 do conjunto de dados e na sequencia vai adicionando os outros elementos utilizando os parâmetros Θ e q. Para adicionar um novo elemento xi é determinado o grupo Ck mais próximo linha (4), isto é, com menor distância d(xi, Tk) . Se esta distância é menor que o limiar Θ o elemento é adicionado nesse grupo Ck. Caso contrário um novo grupo é criado, linha (5)-(7), 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, linha (8).

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

    (1) m = 1
    (2) C1 = {x1}
    (3) T1 = {x1}
    (4) Para i = 2..n
    (5)   Encontrar Ck tal que d(xi, Tk) = min 1≤j≤m d(xi, Tj).
    (6)    Se (d(xi, Tk) > Θ) e (m < q)  então
    (7)        m = m + 1
    (8)        Cm = {xi}
    (9)        Tk = xi
    (10)   Senão
    (11)       Ck = Ck ∪ {xi}
    (12)       Recalcular Tk

Fim
```

Algoritmo 1. Algoritmo BSAS (*Basic Sequential Algorithmic Scheme*)

Nota-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.

Na Figura 5, uma representação com os dados antes e depois do processamento. Em 5-A os dados de entrada antes do agrupamento, no grupo 5-B dados após o

processamento do agrupamento sequencial pelo BSAS (agrupamento realizado através da formula euclidiana). Para o exemplo dado na Figura 5, os parâmetros de entrada utilizados foram, (q = 4) número total de grupos a ser criado e (Θ = 500) cada grupo deve ser formado com a distância do vetor central (este algoritmo usa o primeiro descritor lido como sendo o primeiro centroide, os demais centroides serão criados quando o valor do próximo descritor lido ficar fora do paramentro limar informado) aos outros vetores com o máximo em 500. Os vetores cuja distância entre o vetor central e o vetor de entrada no qual sua métrica de distância ultrapasse o limite informado em Θ, serão agrupados em outro grupo ou em um novo grupo caso seja necessário, desde que não ultrapasse o número limite informado em q.

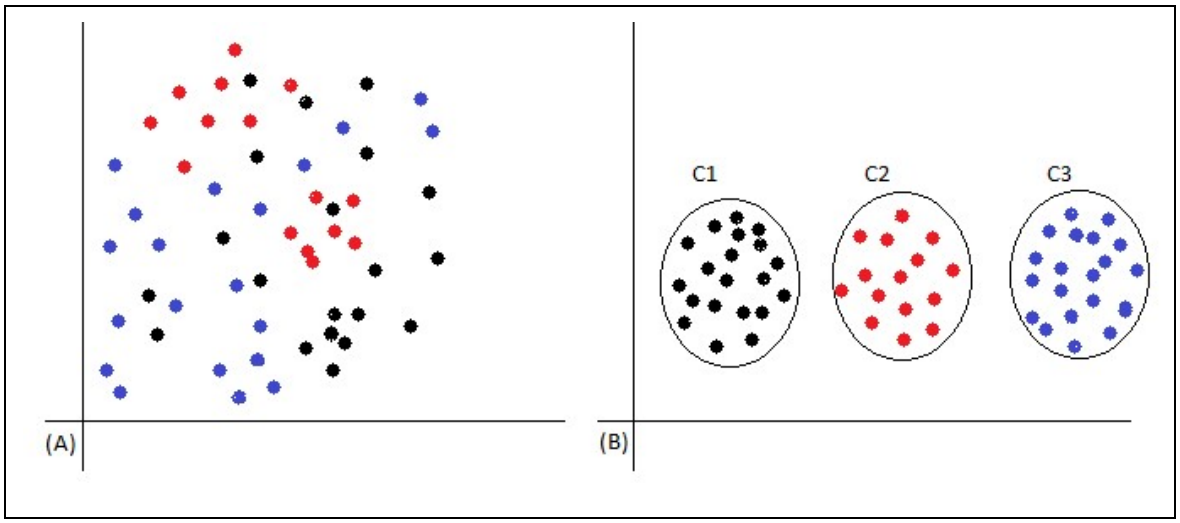

Figura 5. Representação do agrupamento. 5-A dados de entrada sem formação de grupos, 5-B resultado produzido pelo algoritmo BSAS, agrupamento em C_1, C_2 e C_3.

Na representação da Figura 5, tivemos apenas três grupos criados. Para os dados de entrada não ouve a necessidade da criação do quarto grupo passado no parâmetro inicial. Isso indica que nenhum vetor de entrada ficou de fora dos parâmetros passados e três grupos foram suficientes para esse tratamento.

Na Figura 6, um exemplo em que um novo grupo foi gerado, com a chegada de novos vetores para inclusão, seguindo os mesmos parâmetros informados para a formatação da Figura 5, onde (q =4) e (Θ = 500). Os novos descritores não se enquadraram nos grupos já existentes, gerando assim um novo grupo.

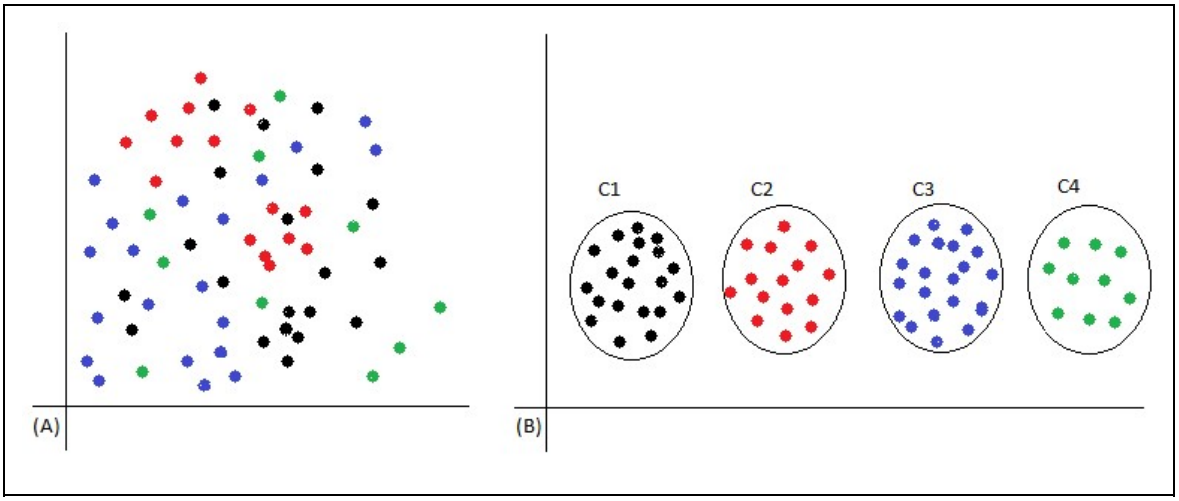

Figura 6. Representação de um novo grupo criado. 6-A dados de entrada com novos vetores a serem agrupados, 5-B resultado do agrupamento com a formação do quarto grupo.

De forma simples o algoritmo gerou um novo grupo sem precisar processar todos os dados anteriores já agrupados, otimizando assim o custo operacional.

4.2. MBSAS - Modified Basic Sequential Algorithmic Scheme

O algoritmo MBSAS (Theodoridis & Koutroumbas 2009) pode ser considerado uma versão melhorada do algoritmo BSAS na qual o processo de formação de grupos é refinado.

Para tanto, em contrapartida, o conjunto de dados deve ser processado duas vezes. O algoritmo 2, na primeira fase linhas (1)-(9) os grupos são definidos dentro dos parâmetros (Θ, q) informados no inicio do processamento e com a incorporação de dados a estes grupos criados. Repetição do Algoritmo 1 BSAS, porém, deixando de agrupar os descritores com similaridade fora do Θ informado no inicio do processamento quando parâmetro q tenha atingido seu limite máximo de grupos. Na segunda fase, os dados que não foram incorporados a qualquer dos grupos durante a primeira fase, linha (10)-(12) voltam a ser processados para identificar o grupo mais similar ao qual possam ser agregados [Real. M. E, 2014]. Na segunda fase não são criados novos grupos, apenas incluídos os descritores que não atenderam ao parâmetro Θ informado pelo usuário e agora serão incorporados ao grupo mais similar a ele.

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

   (1)    m = 1
   (2)    C1 = {x1}
   (3)    Tk = {x1}
   (4)    Para i = 2..n
   (5)      Encontrar Ck tal que d(xi, Tk) = min 1≤j≤m d(xi, Tj).
   (6)      Se (d(xi, Tk) > Θ) e (m < q )  então
   (7)        m = m + 1
   (8)        Cm = {xi}
   (9)        Tk = xi
   (10)   Para i = 2..n
   (11)     Se xi não foi atribuído a nenhum grupo
   (12)       Encontrar Ck tal que d(xi, Tk) = min 1≤j≤m d(xi, Tj).
   (13)       Ck = Ck ∪ {xi}
   (14)       Recalcular T
   (15)   Fim
```

Algoritmo 2. Algoritmo MBSAS (Modified Basic Sequential Algorithmic Scheme).

Como dito no inicio da seção e representado no Algoritmo 2, o algoritmo sequencial MBSAS processa duas vezes os vetores de entrada. Na primeira gera os grupos e agrupa todos que estiverem dentro dos parâmetros passados pelo usuário (Θ, q) e em seguida um segundo processamento agrupa os vetores que ficaram de fora da primeira tentativa.

Na Figura 7, uma representação do agrupamento. As cores mais claras são os vetores que não se enquadraram no primeiro processamento. O grupo (A) representa os vetores de entrada, o grupo (B) representa o primeiro processamento no qual alguns vetores ficaram fora dos grupos criados. Isso ocorreu em virtude dos parâmetros passados pelo usuário, (q = 4) e (Θ = 500). Desta forma os quatro grupos foram criados, porém, alguns vetores não se enquadraram na parametrização atribuída na entrada do processamento. No grupo (C), ocorreu o segundo processamento, verificando apenas os vetores que ficaram de fora dos grupos existentes, associando-os aos grupos mais significativos, embora não o mais representativo.

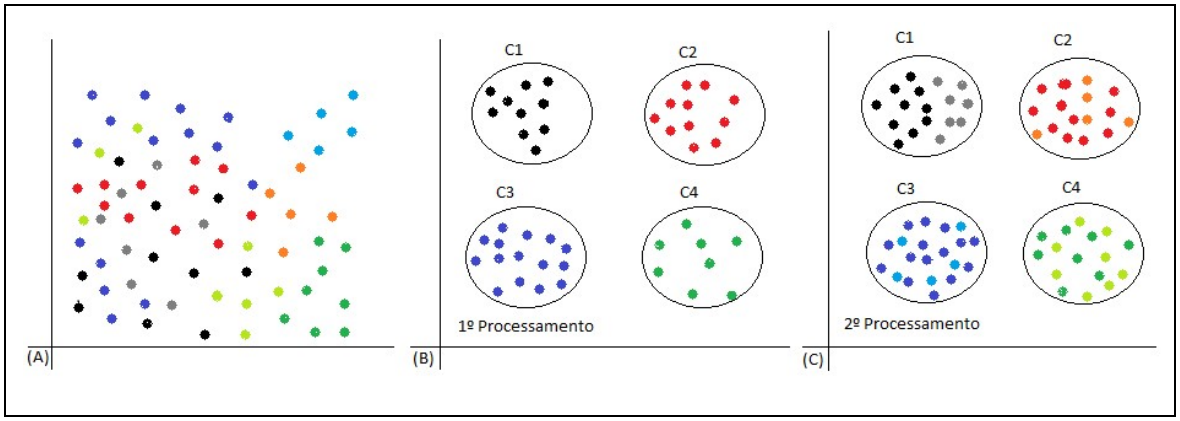

Figura 7. Representação de dois processamentos para os mesmos vetores de entrada.

Capítulo 5. Busca com algoritmos de agrupamento sequencial

Neste capítulo será apresentado o algoritmo de busca eficiente e a forma em que serão demonstrados os resultados na seção 6.5 de experimentos.

5.1. Algoritmo de busca utilizando algoritmos de agrupamento sequencial

Nesta dissertação propomos uma abordagem de busca eficiente utilizando algoritmos de agrupamento sequencial apresentados no capítulo 4. Em particular são utilizados os algoritmos BSAS e MBSAS.

Existem algumas características importantes destes algoritmos que se evidenciam quando consideramos um banco de dados como o conjunto de agrupamentos resultantes destes algoritmos. Estas características são as seguintes:

- Rapidez na criação dos grupos e
- Fácil crescimento do banco de dados, inserção de novos vetores sem precisar reprocessar os dados já agrupados.

Os algoritmos de agrupamento BSAS e MBSAS foram utilizados para organizar em grupos um banco de dados de descritores biométricos faciais. Como função de distância para os algoritmos BSAS e MBSAS foi 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 Xc, um conjunto de agrupamentos C (agrupados pelos algoritmos BSAS e MBSAS) que corresponde ao banco de dados, um conjunto T (também vindo do agrupamento gerado pelos algoritmos BSAS e MBSAS (centroide)) que corresponde aos centróides desses agrupamentos e um parâmetro p para taxa de penetração. A taxa de penetração define qual o percentual de agrupamentos deverão ser explorados na busca

do descritor de consulta Xc. Quando a taxa de penetração for igual a 100 o algoritmo realizará a busca exaustiva percorrendo todos os descritores de todos os agrupamentos no banco de dados.

Como descrito no Algoritmo 3, o algoritmo primeiramente procura os S1,...,Sp agrupamentos mais similares a Xc. Para isto, nas linhas (1)-(3) é construído um vetor D contendo em cada elemento Di a distância de Xc a cada ponto Ti que corresponde ao centróide do agrupamento Ci e um vetor Si auxiliar de agrupamentos. Quanto menor a distância maior será a similaridade. Na sequência, na linha (4), a lista de agrupamentos Si é ordenada de maneira ascendente segundo o vetor Di. Desta maneira os p primeiros agrupamentos em S serão os mais similares a Xc. A seguir, nas linhas (7) – (11) é calculada a similaridade de Xc com cada um dos descritores nesse conjunto de agrupamentos S1,.,Sp para encontrar e retornar o descritor mais similar minX.

```
Algoritmo BuscaEficiente
 Entrada: Descritor de consulta: Xc
    Conjunto de Agrupamentos: C = { C1 , C2 , ....., Ck }
    Conjunto de Centroídes: T =  { T1 , T2 , ....., Tk }
    Função de distância: dist
    Taxa de penetração: p

 Saída: Descritor mais similar a xc em C : minX

 Auxiliar: Lista de distancias: D
           Lista de Agrupamentos: S
      (1) Para cada Ci ∈ C
      (2)    Di =  dist(xc , Ti))
      (3)    Si = Ci;
      (4) OrdenarDistancias(Si, Di)
      (5) minD = ∞
      (6) minX = Ø
      (7) Para i ∈ [1..p]
      (8)    Para cada xj ∈ Si
      (9)       Se  dist(xc, xj) < minD
      (10)         minD = d(xc, xj)
      (11)         minX = xj
      (12) Retornar minX
```

Algoritmo 3. Algoritmo de busca eficeinte em grupo.

Se analisarmos o algoritmo de busca verificamos que o custo computacional ou número de comparações da busca pode ser descrito pela fórmula

O (n + nlnn + Som(i= 1, p , C(i)))

Sendo que:

n : número de agrupamentos

p: taxa de penetração

C (i) : quantidade de descritores no agrupamento i

Nota-se que o termo Som(i= 1, p , C(i)) descreve a quantidade de descritores no banco de dados que efetivamente foram comparados. A este termo chamamos de Taxa de penetração no banco de dados d = Som(i= 1, p , C(i))

O (n + nlnn + d))

Sendo que

n : número de agrupamentos

p: taxa de penetração

d: taxa de penetração no banco de dados

Um banco de dados ideal seria aquele com uma distribuição uniforme dos descritores nos agrupamentos, isto é onde os valores de todos os C(i) sejam iguais.

O algoritmo *BuscaEeficiente* foi desenvolvido para avaliar os dados organizados pelas técnicas de agrupamento sequêncial. A finalidade desta avaliação é encontrar o menor percentual de penetração no banco de dados com a mesma eficiência como se fosse percorrido o banco de dados por completo. Isso de forma que a margem de erro seja aceitável e com o resultado mais próximo possível de quando realizado a busca de forma exaustiva.

5.2. Apresentação de resultados

Como mencionado na introdução, um problema de metodologia encontrada nos trabalhos correlatos é o fato de que a precisão é calculada para taxas de penetração nos grupos. Nota-se que a taxa de penetração nos grupos não necessariamente corresponde à mesma taxa de penetração no banco de dados. A busca de um descritor sempre é realizada em uma quantidade fixa de grupos mais similares. No entanto, como a quantidade de descritores em cada grupo não é uniforme, a busca de um descritor pode envolver a seleção de grupos diferentes e, portanto, a quantidade de descritores comparados pode ser também diferente. Por esta razão, para medir qual o percentual do banco de dados explorado para cada valor de taxa de penetração nos grupos, utilizamos o percentual médio de descritores percorridos e o desvio padrão da quantidade de descritores comparados.

No Capítulo 6, serão apresentadas as evidencias obtidas pelos testes realizados. No exemplo dado abaixo na Figura 8, o gráfico mostra nas abscissas a variação da taxa de penetração no agrupamento.

O gráfico superior mostra para cada valor de taxa de penetração no agrupamento qual a taxa de precisão, isto é, o percentual de consultas que obtiveram o mesmo resultado que a busca exaustiva, isto é, que foram corretamente recuperados.

O gráfico inferior mostra qual a média e variância da quantidade de comparações de descritores no banco de dados, isto é, a média e variância da taxa de penetração no banco de dados.

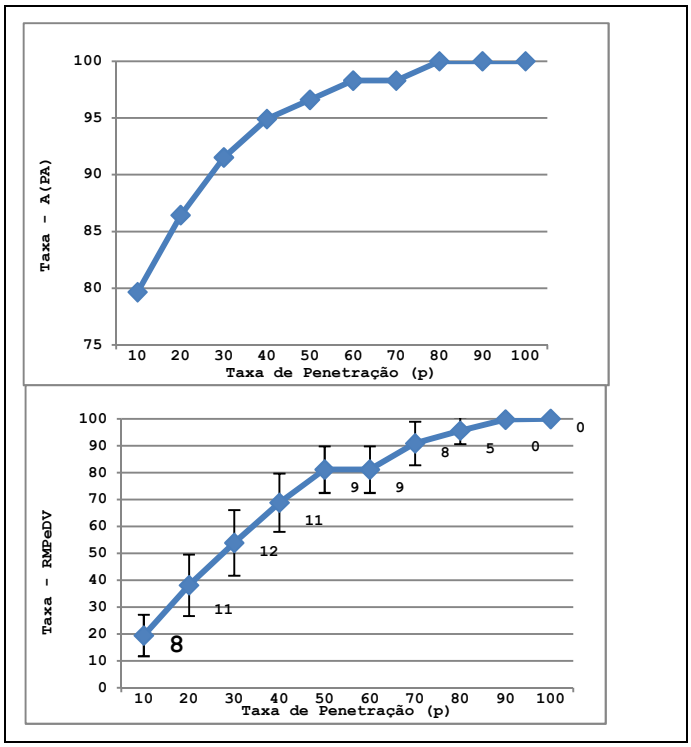

Figura 8. Exemplo de demonstração dos dados evidenciados (Capítulo 6).

Nesta representação são mostrados simultaneamente a taxa de penetração dos grupos e a taxa de penetração no banco de dados, desta forma é possível avaliarmos a taxa de acerto e também a quantidade de registro percorridos já que os grupos são formados por similaridade. Pelo fato do agrupamento ocorrer por similaridade, a probabilidade de dois ou mais grupos terem a mesma quantidade de descritores é mínima ou até mesmo nula.

Estas informações adicionais nos permitem visualizar além da taxa de acerto, as variações de descritores acessados para cada parâmetro de taxa de penetração passado ao algoritmo.

Olhando para Figura 8, Gráfico inferior, na linha Taxa de Penetração, os valores de 10 a 100 representam o percentual de penetração nos grupos criados no banco de dados a partir dos algoritmos BSAS e MBSAS. Na linha Ponto Médio e Desvio Padrão, por exemplopara Taxa de Penetração igual a 10, foram percorridos 20% dos descritores e teve um Desvio Padrão de 8 descritores, para mais ou para menos. Esses 10% no grafico inferior representam um acerto de 80% dos descritores pesquisados em apenas 20% do banco de dados percorrido, representando uma redução no custo da busca.

Capítulo 6. Experimentos e Análise dos Resultados

Para avaliar o algoritmo de busca eficiente proposto, foi utilizado um Banco de Dados formado por um conjunto de descritores biométricos de teste e um conjunto de descritores de consulta. O banco de dados de teste é formado por um conjunto de 2.600 descritores da face de 1.093 indivíduos diferentes obtidos de imagens faciais do banco de dados FERET [Phillips, P. et. al. 2000]. O conjunto de descritores de consulta é formado por 165 descritores, que não fazem parte do banco de dados, mas que estão associados a indivíduos que possuem algum descritor armazenado no banco de dados de testes. Estes descritores foram obtidos pela aplicação do método EigenFaces [Turk & Pentland 1991] em cada uma das imagens faciais. A função de similaridade utilizada sobre estes descritores foi à distância euclidiana.

Para a criação dos agrupamentos através dos algoritmos BSAS e MBSAS, os descritores foram incluídos de forma aleatória. O impacto de diferentes ordenações dos descritores durante a criação dos agrupamentos não será abordado neste trabalho.

Para avaliar o algoritmo é calculada a precisão quando o espaço de busca é reduzido segundo uma determinada taxa de penetração. Definimos como precisão a quantidade de descritores de consulta para os quais se consegue encontrar o mesmo descritor que na busca exaustiva.

Em uma consulta em um banco de dados biométrico, nem sempre se garante que o descritor de maior similaridade encontrado corresponde a um descritor do mesmo indivíduo. Esta taxa de erro está associada à precisão do algoritmo biométrico que gera esses descritores. Por esta razão, neste trabalho selecionamos para o estudo unicamente os descritores de consulta para os quais, na busca exaustiva, é encontrado um descritor do mesmo individuo.

Os experimentos foram realizados com a variação dos parâmetros Θ e q dos algoritmos de agrupamento sequencial BSAS e MBSAS. Em cada experimento foi realizada a variação da taxa de penetração e para cada uma delas foi determinada o percentual de descritores médios percorridos e o desvio padrão da quantidade de descritores (RMP e DV) e a taxa de precisão A(PA).

6.1. Metodologia

A pesquisa visa avaliar os algoritmos sequencias BSAS e MBSAS propostos aqui. Esta avaliação pretende demonstrar a eficiência destes algoritmos de agrupamento ao fato de reduzir o acesso ao banco de dados em tempo de pesquisa e com resultado eficiente na mesma proporção realizada pela busca exaustiva. Este processo de avaliação não está amarrado a um tipo especifico de banco de dados, qualquer banco biométrico que tenha seus descritores agrupados em grupos pode ser medido pelo algoritmo de busca por redução de espaço apresentado no capítulo 5.

A abordagem utilizada pretende medir a precisão de acerto da busca, a quantidade média de descritores percorridos e o desvio padrão. Este processo será medido conforme for estreitando o campo de busca, ou seja, com o auxilio do parâmetro taxa de penetração, atribuindo a ele um percentual de grupos a serem percorridos. Através deste percentual o algoritmo calcula a quantidade de grupos que devem ser percorrido para realizar o reconhecimento do individuo de busca.

6.2. Parâmetros avaliados

Os experimentos realizados incluíram a avaliação da precisão de acerto, quantidade média de descritores percorridos e desvio padrão com variação dos seguintes parâmetros:

1. Algoritmos de agrupamento (BSAS e MBSAS).
    a. Quantidade máxima (grupos a serem gerados).
    b. Limiar máximo (distância entre dois vetores).

2. Algoritmo de busca.
    a. Taxa de penetração.
    b. Descritores Médios percorridos.
    c. Desvio Padrão.

3. Método de similaridade.
    a. Função euclidiana.

6.3. Implementações realizadas

Para realizar os experimentos foram realizadas as seguintes implementações:

1. Programa de carga inicial, no qual gera uma base de indivíduos para consultas de forma aleatória.
2. Programa criado com base no pseudocódigo BSAS para agrupamento dos descritores.
3. Programa criado com base no pseudocódigo MBSAS para agrupamento dos descritores.
4. Programa criado para Busca Eficiente.

Os algoritmos foram implementados na linguagem COBOL em processos Batch e com auxilio do banco de dados DB2.

6.4. Experimentos realizados

Para calcular a eficiência do método, inicialmente é realizada uma busca exaustiva no banco de dados, selecionando os descritores mais similares aos 165 descritores de consulta e comparando se o mais similar encontrado é do mesmo individuo de consulta. Quando realizamos uma consulta em um banco de dados biométrico, nem sempre o descritor de maior similaridade encontrado corresponde a um descritor do mesmo indivíduo. Este problema é consequência da precisão do método biométrico que gera esses descritores.

Nota-se que a busca exaustiva corresponde a uma taxa de penetração de 100% dos descritores contidos no banco de dados. Quando a taxa de penetração é reduzida, calculamos a precisão determinando a quantidade de descritores de consulta que conseguem encontrar o mesmo descritor da busca exaustiva. Uma taxa de penetração do conjunto de agrupamentos não necessariamente corresponde à taxa de penetração no banco de dados. Isto acontece porque a quantidade de descritores em cada agrupamento pode ser diferente. Desta maneira, como parte dos experimentos, foi calculado o percentual do banco de dados que precisou realmente ser comparado para uma taxa de penetração determinada. Como cada busca de um descritor pode envolver a seleção de agrupamentos diferentes, a quantidade de descritores comparados pode ser também

diferente. Para medir o percentual do banco de dados explorado utilizamos a média dos descritores percorridos e o desvio padrão da quantidade de descritores comparados.

Para os experimentos, foram medidas a taxa de precisão de agrupamento A(PA), a taxa do percentual de descritores médios percorridos e desvio padrão (RMP e DV) quando modificada a taxa de penetração (p) do algoritmo de *busca por redução de espaço*. Os resultados dos experimentos apresentados foram obtidos com a variação dos parâmetros Θ e q dos algoritmos de agrupamento sequencial BSAS e MBSAS.

Três parâmetros foram avaliados para cada parâmetro de busca realizado (taxa de penetração no banco de dados):

1. Taxa de precisão de acerto A(PA)
    a. Mede o percentual de acerto para cada taxa de penetração informada no parâmetro de busca do algoritmo busca por bedução de espaço.
2. Percentual de descritores/descritores médios percorridos
    a. Mede a quantidade e o percentual médio de descritores percorridos para cada taxa de penetração informada.
3. Desvio padrão
    a. Mede o desvio padrão sobre os descritores médios percorridos do item 2.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2},$$

Para o agrupamento realizado pelos algoritmos BSAS e MBSAS, serão passados os parâmetros Θ (limear) e q (quantidade máxima de grupos a serem criados) como mencionados nos capítulos anteriores. Para cada Θ teremos uma variação em q na qual serão gerados quatro agrupamentos diferentes. Para cada agrupamento realizado executaremos o algoritmo *de busca por redução de espaço* e com base no retorno serão geradas as evidências conforme descrido na Seção 5.1.

Nas Seções 6.5 e 6.6 serão apresentadas tabelas com os gráficos correspondentes às buscas realizadas e nos anexos a planilha com os dados que geraram os gráficos destas tabelas e outros evidências que foram coletadas mas que não foram inseridas aqui e estão nos anexos para simples conferência ou validação.

Para cada tabela de evidências exibida nas próximas seções os valores do parâmetro q sempre serão os mesmos (05, 08, 12 e 20) o que vai variar de uma tabela para outra é valor do Θ.

6.5. Resultado obtido pelo agrupamento do algoritmo BSAS.

Nesta seção serão apresentados os resultados do algoritmo BSAS na medida em que aumenta o limiar de distância e variação da quantidade de grupos criados.

A Figura 9 apresenta os resultados obtidos pelo algoritmo de busca por redução de espaço para descritores agrupados segundo o algoritmo BSAS. Estes resultados foram obtidos para variação do parâmetro de limiar de distância Θ com valor 100.

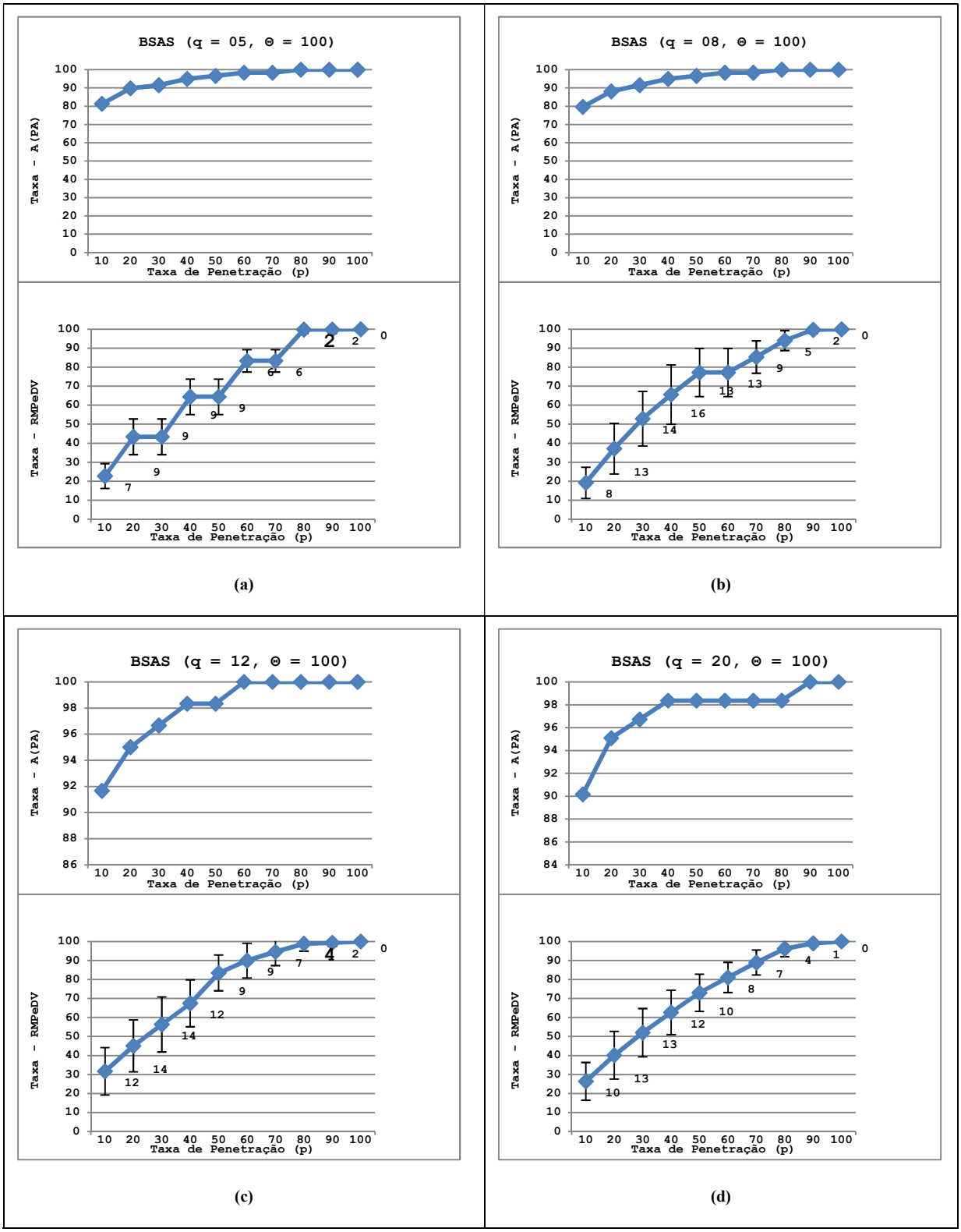

Figura 9. Resultados da busca reduzida pelo agrupamento do algoritmo BSAS no parâmetro (Θ) com valor 100 e com variação de q (a-05, b-08, c-12 e d-20).

Tabela 1. Melhor resultado de cada agrupamento na menor taxa penetração no parâmetro (Θ*)* com valor 100 e com variação de q (05, 08, 12 e 20).

Resultado BSAS				
Grupo /Limiar	**Percentual Taxa de penetração**	**Percentual Taxa de acerto**	**Percentual médio de Descritores percorridos na busca**	**Percentual Desvio padrão sobre o Percentual médio**
05/100	80	100	99,8	2
08/100	80	100	94,1	5,3
12/100	60	100	90,3	9,1
20/100	90	100	99,1	1,5

O melhor resultado foi alcançado na figura 9-c com um valor q igual a 12 e a taxa de penetração igual a 60. Este experimento mostrou uma redução de 10% referente a comparação do descritor de busca com os descritores contidos no banco de dados. Figuras 9-a e 9-b atingiram 100% de acerto em 80% da penetração. Olhando para a Tabela 1, onde os dados então mais claros e originaram o gráfico da Figura 9, podemos ver com clareza que em tabela 1 ((05/100) e (08/100)) na mesma taxa de penetração houve mudança no percentual médio de descritores percorridos, Tabela 1 (05/100) 99,8 e em tabela 1 (08/100) 94,1.

Observando os dois melhores resultados das figuras 9-b e 9-c, a variação de descritores percorridos foi de 4,1% no pior cenário, que é um bom resultado para esse tipo de consulta. Porém, olhando pela taxa de penetração, figura 9-c atingiu o valor máximo de acerto com apenas 60% e isso mostra o quanto é importante a medição do percentual médio de descritores percorridos e não apenas a taxa de penetração. A diferença da taxa de penetração foi relativamente grande já a quantidade de descritores percorridos nem tanto. Figura 9-d teve o pior nível de acerto, a taxa de penetração foi de 90% e 99,3% de descritores percorridos no banco de dados.

A Figura 10 mostra os resultados que foram obtidos pela variação do parâmetro de limiar de distância Θ com valor 300.

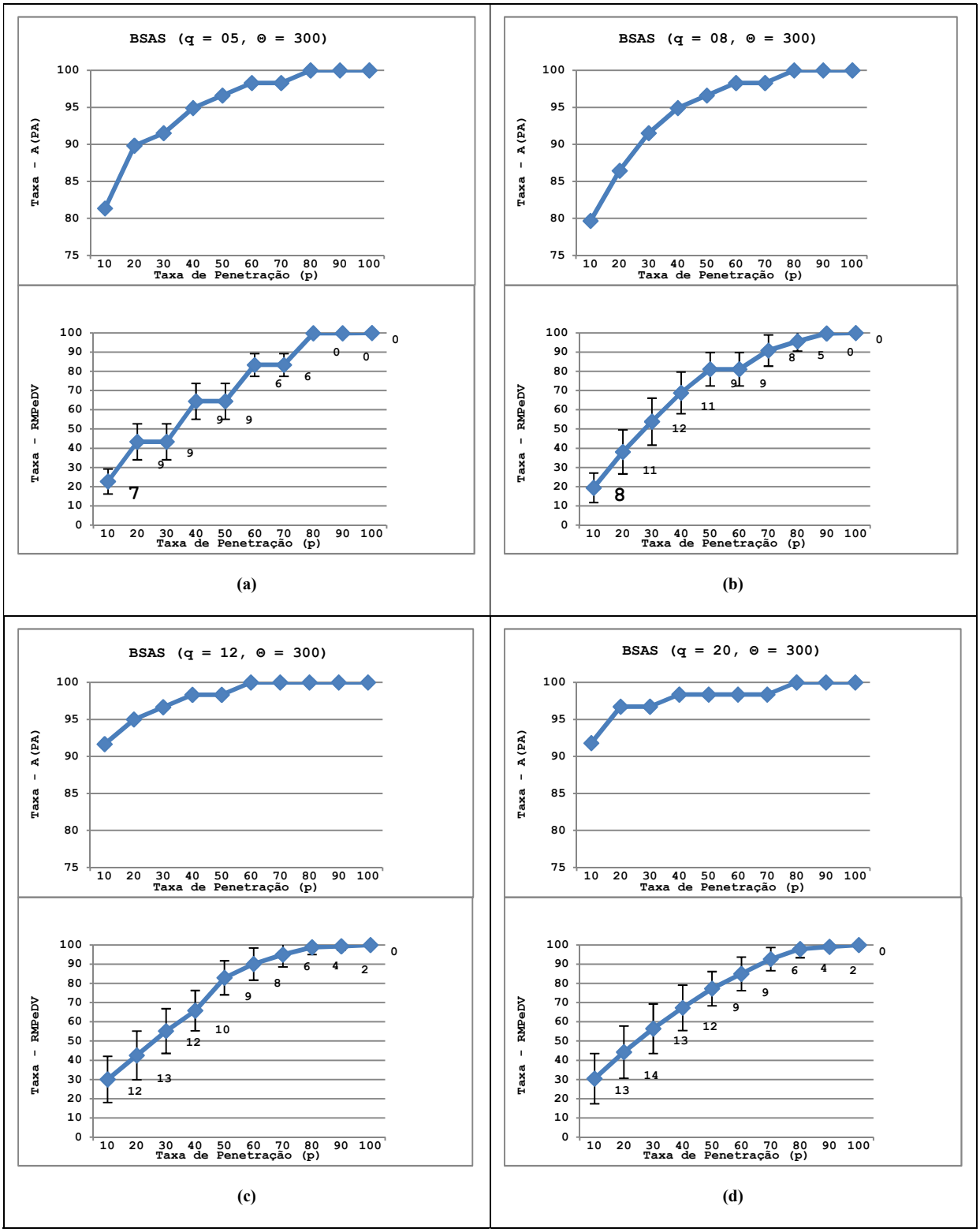

Figura 10. Resultados da busca reduzida pelo agrupamento do algoritmo BSAS no parâmetro (Θ) com valor 300 e com variação de q (a-05, b-08, c-12 e d-20).

Tabela 2. Melhor resultado de cada agrupamento na menor taxa penetração no parâmetro (*Θ)* com valor 300 e com variação de q (05, 08, 12 e 20).

Resultado BSAS				
Grupo/Limiar	**Percentual Taxa de penetração**	**Percentual Taxa de acerto**	**Percentual médio de Descritores percorridos na busca**	**Percentual Desvio padrão sobre o Percentual médio**
05/300	80	100	99,8	2
08/300	80	100	95,6	5
12/300	60	100	90,3	8,4
20/300	80	100	97,9	4,5

Este experimento teve o mesmo ponto de equilíbrio atingido na figura 9-c, onde o melhor resultado foi alcançado com a taxa de penetração em 60 e com um valor q igual a 12. A alteração em relação a figura 9-c foi no percentual médio de descritores percorridos representados em tabela 2 (12/300), que foi de 90,3. Aumento pouco representativo, pois a variação foi de apenas 0,1%.

Nas figuras 10-a, 10-b e 10-d atingiram 100% de acerto com a taxa de penetração em 80. As alterações entre elas ficaram por conta do percentual médio de descritores percorridos. Figura 10-a (tabela 2 – 05/300) com 99,8, figura 10-b (tabela 2 – 08/300), com 95,6 e figura 10-d (tabela 2 – 20/300), com 97,9.

Os dois melhores resultados foram nas figuras 10-b e 10-c, onde a variação de descritores percorridos foi de 5,1% no pior cenário.

Na Figura 11, os resultados foram obtidos para variação do parâmetro de limiar de distância Θ com valor 500.

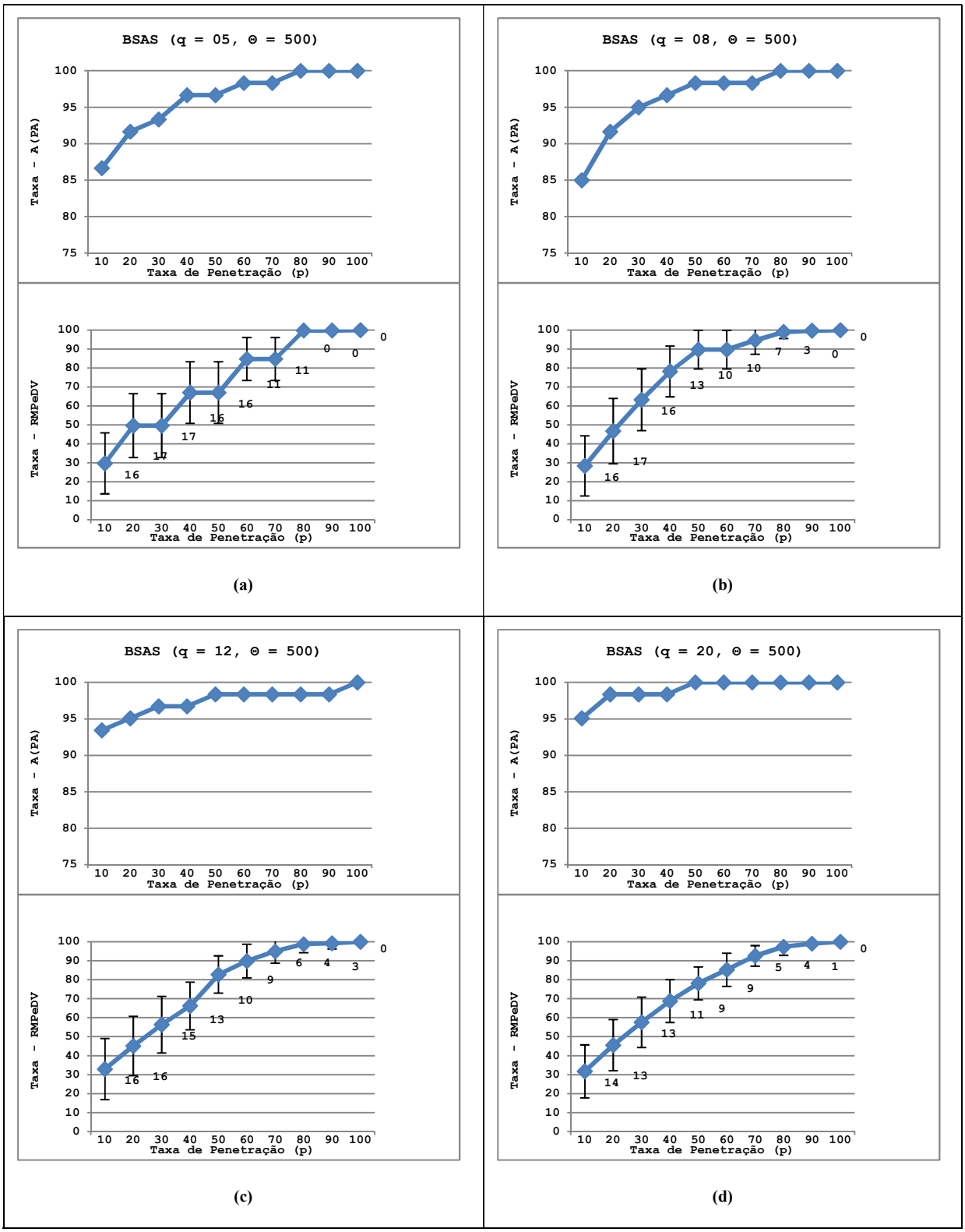

Figura 11. Resultados da busca reduzida pelo agrupamento do algoritmo BSAS no parâmetro (Θ) com valor 500 e com variação de q (a-05, b-08, c-12 e d-20).

Tabela 3. Melhor resultado de cada agrupamento na menor taxa penetração no parâmetro (*Θ)* com valor 500 e com variação de q (05, 08, 12 e 20).

Resultado BSAS				
Grupo/Limiar	**Percentual Taxa de penetração**	**Percentual Taxa de acerto**	**Percentual médio de Descritores percorridos na busca**	**Desvio padrão sobre o Percentual médio**
05/500	80	100	99,8	2
08/500	80	100	99,1	3,4
12/500	100	100	100	0
20/500	50	100	78,6	8,6

Este experimento teve melhor ponto de equilíbrio em figura 11-d, com o melhor resultado alcançado com a taxa de penetração em 50, com valor de q igual a 20 e percentual médio de descritores percorridos na casa de 78,6%.

Nas figuras 11-a e 11-b, com a taxa de penetração em 80, atingiram o mesmo valor de acertos da busca exaustiva e, com variação dos descritores percorridos, com uma diferença de apenas 0,7%. Figura 11-a (tabela 3 – 05/500) com 99,8, figura 11-b (tabela3 – 08/500), com 99,1. Figura 11-c teve o pior desempenho, nenhumas das reduções realizadas com a taxa de penetração conseguiram atingir o mesmo resultado da busca exaustiva.

A Figura 12, os resultados foram obtidos para variação do parâmetro de limiar de distância Θ com valor 1000.

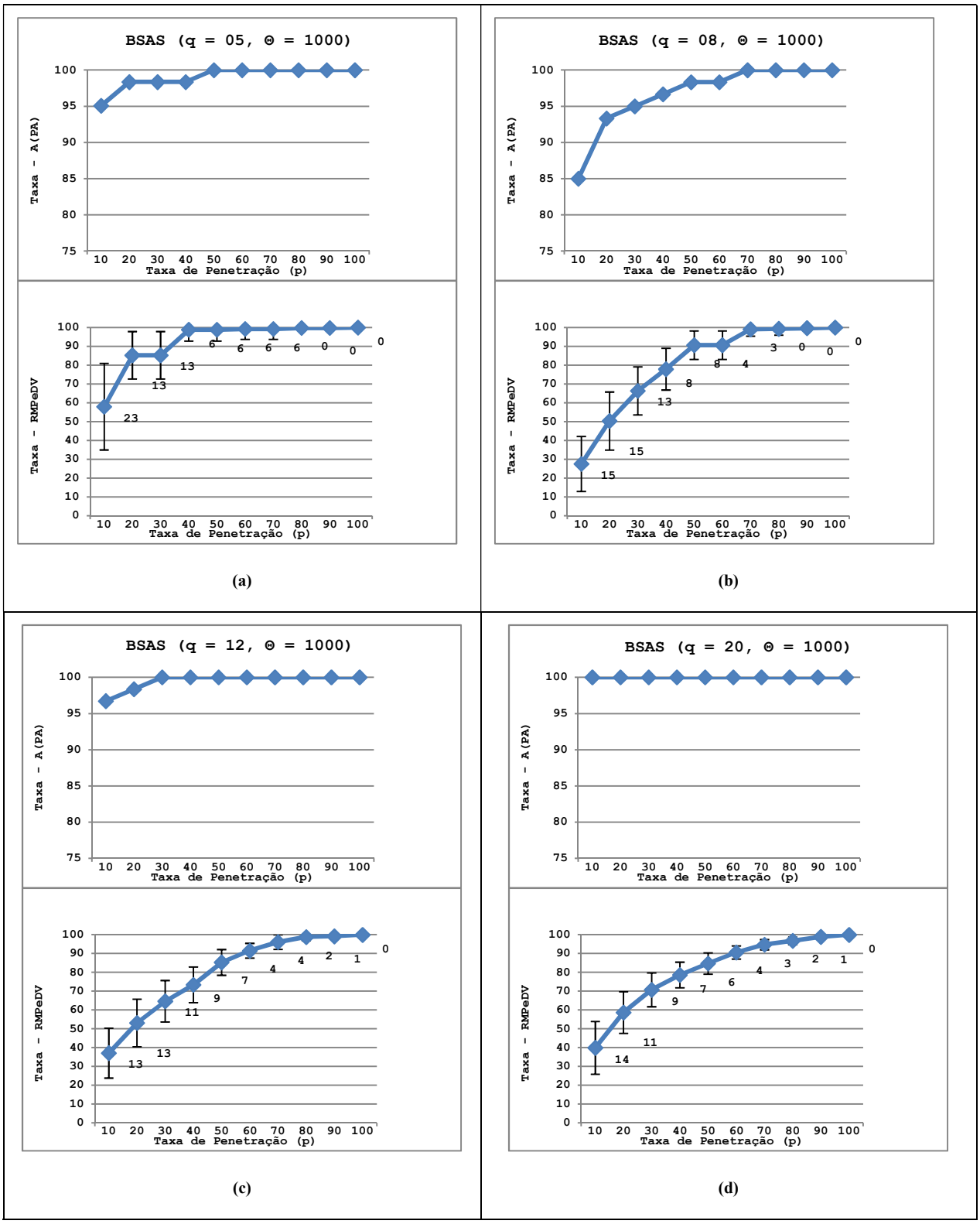

Figura 12. Resultados da busca reduzida pelo agrupamento do algoritmo BSAS no parâmetro (Θ) com valor 1000 e com variação de q (a-05, b-08, c-12 e d-20).

Tabela 4. Melhor resultado de cada agrupamento na menor taxa penetração no parâmetro (Θ) com valor 1000 e com variação de q (05, 08, 12 e 20).

Resultado BSAS				
Grupo/Limiar	**Percentual Taxa de penetração**	**Percentual Taxa de acerto**	**Percentual médio de Descritores percorridos na busca**	**Percentual Desvio padrão sobre o Percentual médio**
05/1000	50	100	98,9	6,1
08/1000	70	100	99,1	3,6
12/1000	30	100	64,7	11,1
20/1000	10	100	40	14

Os experimentos da figura 12 apresentaram excelentes resultados para a taxa de penetração. O pior resultado foi para figura 12-b (tabela 4 – 08/1000) com taxa de penetração em 70 e seu percentual médio de descritores percorridos em 99,1. O melhor resultado foi obtido em figura 12-d (tabela 4 – 20/1000) onde o melhor resultado alcançado foi com taxa de penetração igual a 10 que é o parâmetro mínimo passado neste experimento e percorreu apenas 40% dos descritores. Outro resultado de boa qualidade se deu na figura 12-c (tabela 4 – 12/100) com taxa de penetração igual a 30 e percentual médio de descritores percorridos em 64,7.

A Tabela 5 apresenta dois resultados, o pior e o melhor resultado gerado para o agrupamento realizado com o algoritmo sequencial BSAS.

Tabela 5. Pior e melhor resultado para o agrupamento realizado pelo algoritmo BSAS.

Resultado BSAS					
Grupo/Limiar	**Percentual Taxa de penetração**	**Percentual Taxa de acerto**	**Percentual médio de Descritores percorridos na busca**	**Percentual Desvio padrão sobre o Percentual médio**	
12/500	100	100	100	0	Pior resultado
20/1000	10	100	40	14	Melhor resultado

A Tabela 5 nos mostra duas situações de respostas do algoritmo de agrupamento BSAS com discrepância dos mesmos. Em tabela 10 (12/500) nenhum taxa de penetração menor que a busca exaustiva ou busca em 100% do banco de dados tiveram bons resultados. Enquanto que em tabela 10 (20/1000) o resultado foi excelente, com apenas 10% de penetração e 40% dos descritores percorridos atingiu o mesmo valor da busca exaustiva. Os parâmetros passados ao método de agrupamento BSAS em tabela 10 (20/1000) gerou uma redução de 60 % no tempo de busca.

Os resultados obtidos mostram que na medida em que aumenta o limiar de distância (Θ) precisamos de uma menor taxa de penetração para atingir valores altos de precisão. Também podemos ver que para conseguirmos uma consistência em valores altos de precisão é necessário um aumento da quantidade de grupos (q). Por outro lado, este aumento da quantidade de grupos em geral aumenta a média da quantidade de descritores que precisam ser comparados para uma mesma taxa de penetração. Em geral, os resultados com o algoritmo BSAS mostram a necessidade de uma alta taxa de penetração para garantir uma precisão de 100%.

6.6. Resultado obtido pelo agrupamento do algoritmo MBSAS.

A Figura 13 mostra os resultados obtidos pelo algoritmo de busca por redução de espaço nos dados agrupados pelo algoritmo de agrupamento sequencial MBSAS com Θ (100).

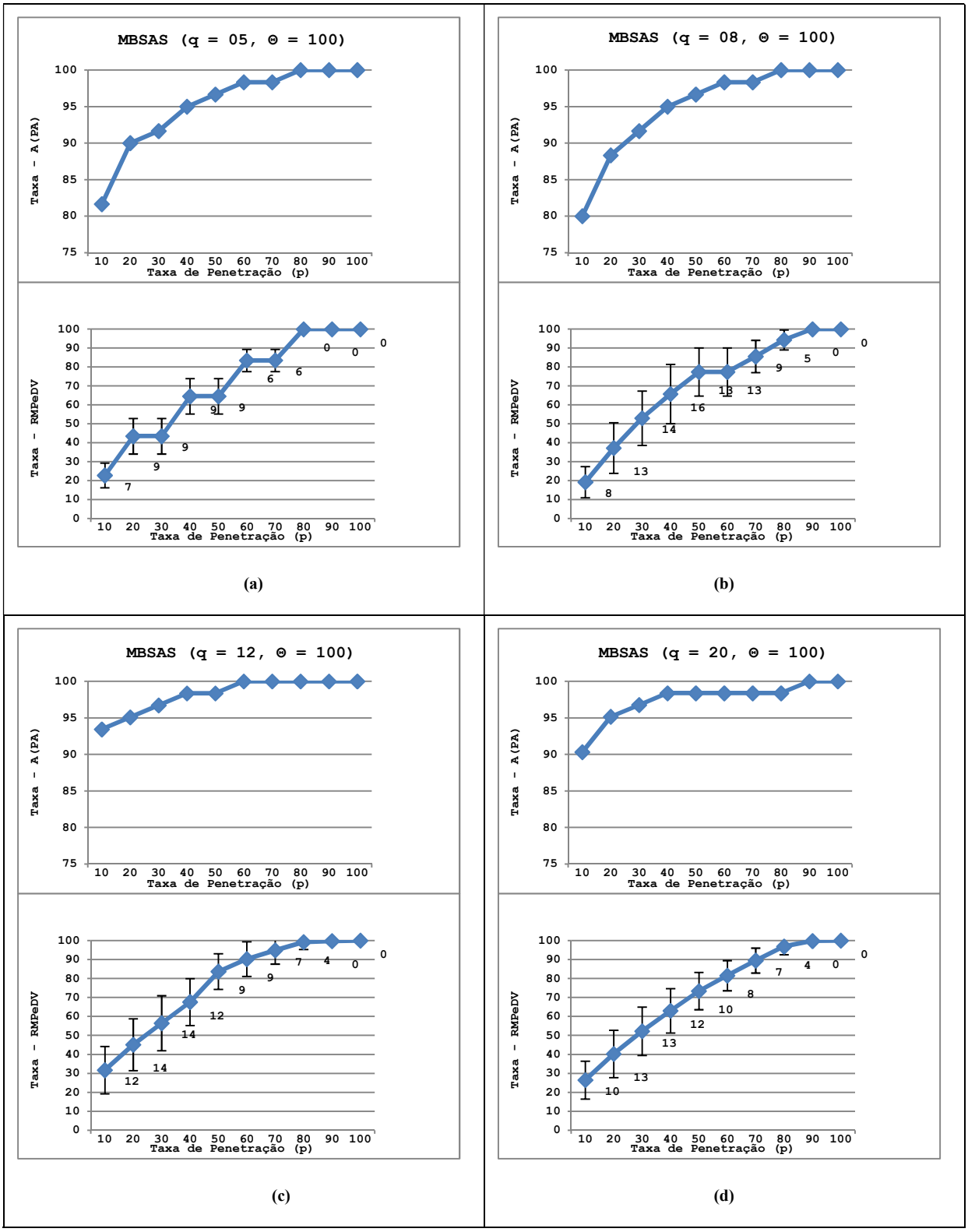

Figura 13. Resultados da busca reduzida pelo agrupamento do algoritmo MBSAS no parâmetro (Θ*)* com valor 100 e com variação de q (a-05, b-08, c-12 e d-20).

Tabela 6. Melhor resultado de cada agrupamento na menor taxa penetração no parâmetro (Θ*)* com valor 100 e com variação de q (05, 08, 12 e 20).

Resultado MBSAS				
Grupo/Limiar	**Percentual Taxa de penetração**	**Percentual Taxa de acerto**	**Percentual médio de Descritores percorridos na busca**	**Desvio padrão sobre o Percentual médio**
05/100	80	100	100	0
08/100	80	100	94,4	5,3
12/100	60	100	90	9,1
20/100	90	100	99,8	1,5

Na Figura 13, o resultado obtido pelo algoritmo de busca por redução de espaço para descritores agrupados segundo o algoritmo MBSAS. Estes resultados foram obtidos pela variação do parâmetro de limiar de distância Θ com valor 100.

O melhor resultado é alcançado na figura 13-c com valor q igual a 12 e a taxa de penetração igual a 60, o qual conseguiu 100% de precisão e percorreu 90% dos descritores contidos no banco de dados. Na sequencia outro resultado bem próximo da figura 13-c foi a figura 13-b, com penetração igual a 80 e percentual médio de descritores percorridos em 94,4, diferença de 4,1% do acesso ao banco de dados. Apesar da figura13-d ter uma taxa de penetração maior que figura 13-a, o percentual médio de descritores percorridos foi 0,1% abaixo da F13-a.

Na Figura 14, os resultados obtido pelo algoritmo de busca por redução de espaço para descritores agrupados segundo o algoritmo MBSAS. Estes resultados foram obtidos pela variação do parâmetro de limiar de distância Θ com valor 300.

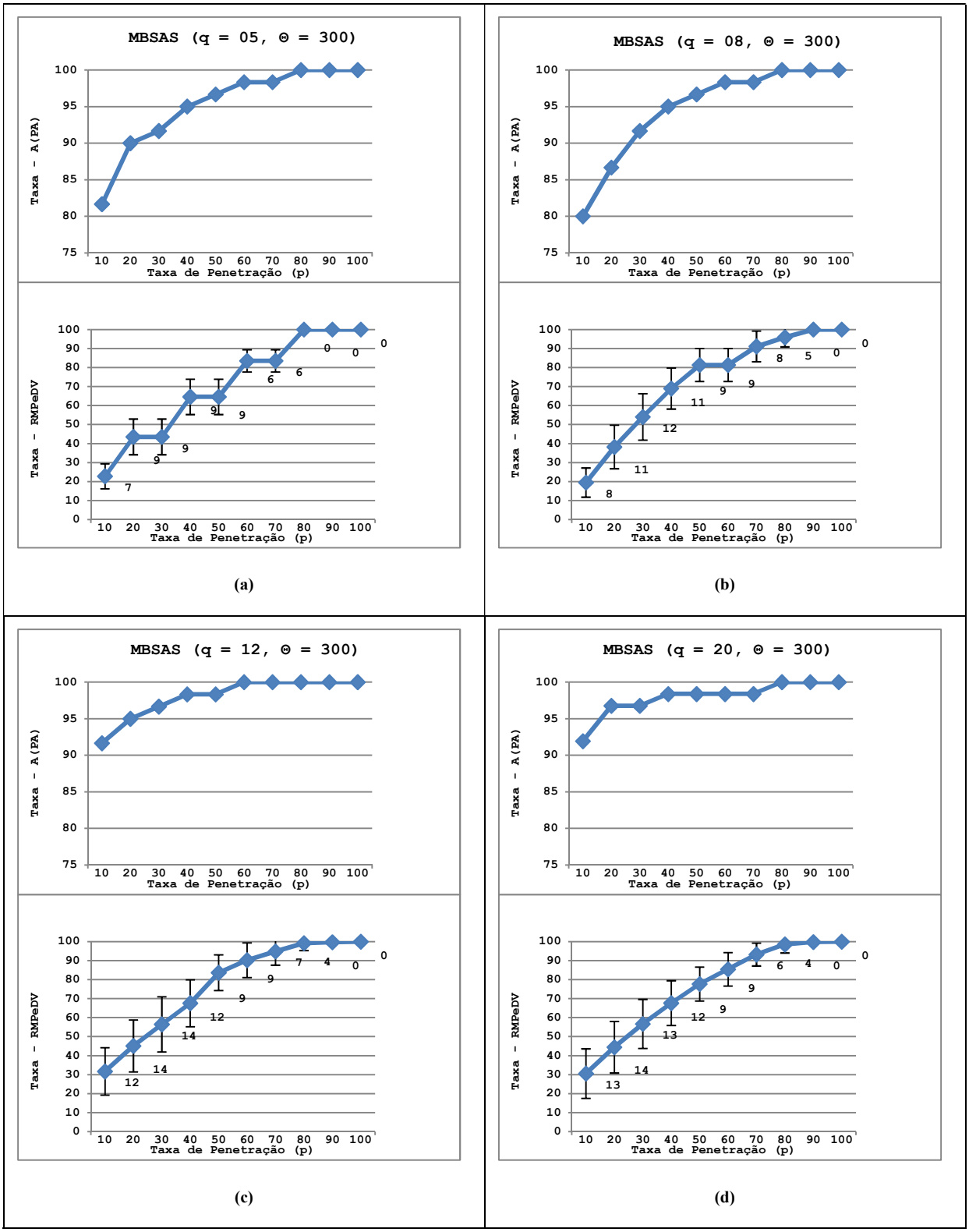

Figura 14. Resultados da busca reduzida pelo agrupamento do algoritmo MBSAS no parâmetro (Θ*)* com valor 300 e com variação de q (a-05, b-08, c-12 e d-20).

Tabela 7. Melhor resultado de cada agrupamento na menor taxa penetração no parâmetro (Θ*)* com valor 300 e com variação de q (05, 08, 12 e 20).

Resultado MBSAS				
Grupo/Limiar	**Percentual Taxa de penetração**	**Percentual Taxa de acerto**	**Percentual médio de Descritores percorridos na busca**	**Desvio padrão sobre o Percentual médio**
05/300	80	100	100	0
08/300	80	100	95,9	5
12/300	60	100	90,1	9,1
20/300	80	100	98,6	4,5

O melhor resultado é alcançado na figura 14-c com valor q igual a 12 e a taxa de penetração igual a 60, o qual conseguiu 100% de precisão e percorreu 90,1% dos descritores contidos no banco de dados, o mesmo resultado da figura 13-c. As demais fiduras 14-a, 14b e 14-d, todas tiveram a mesma taxa de penetração com o mesmo resultado da busca exaustiva, onde essa taxa foi igual a 80, dentre elas o melhor resultado com menor percentual médio de descritores percorridos foi a figura 14-b.

Na Figura 15, os resultados obtidos pelo algoritmo de busca por redução de espaço para descritores agrupados segundo o algoritmo MBSAS, para variação do parâmetro de limiar de distância Θ com valor 500.

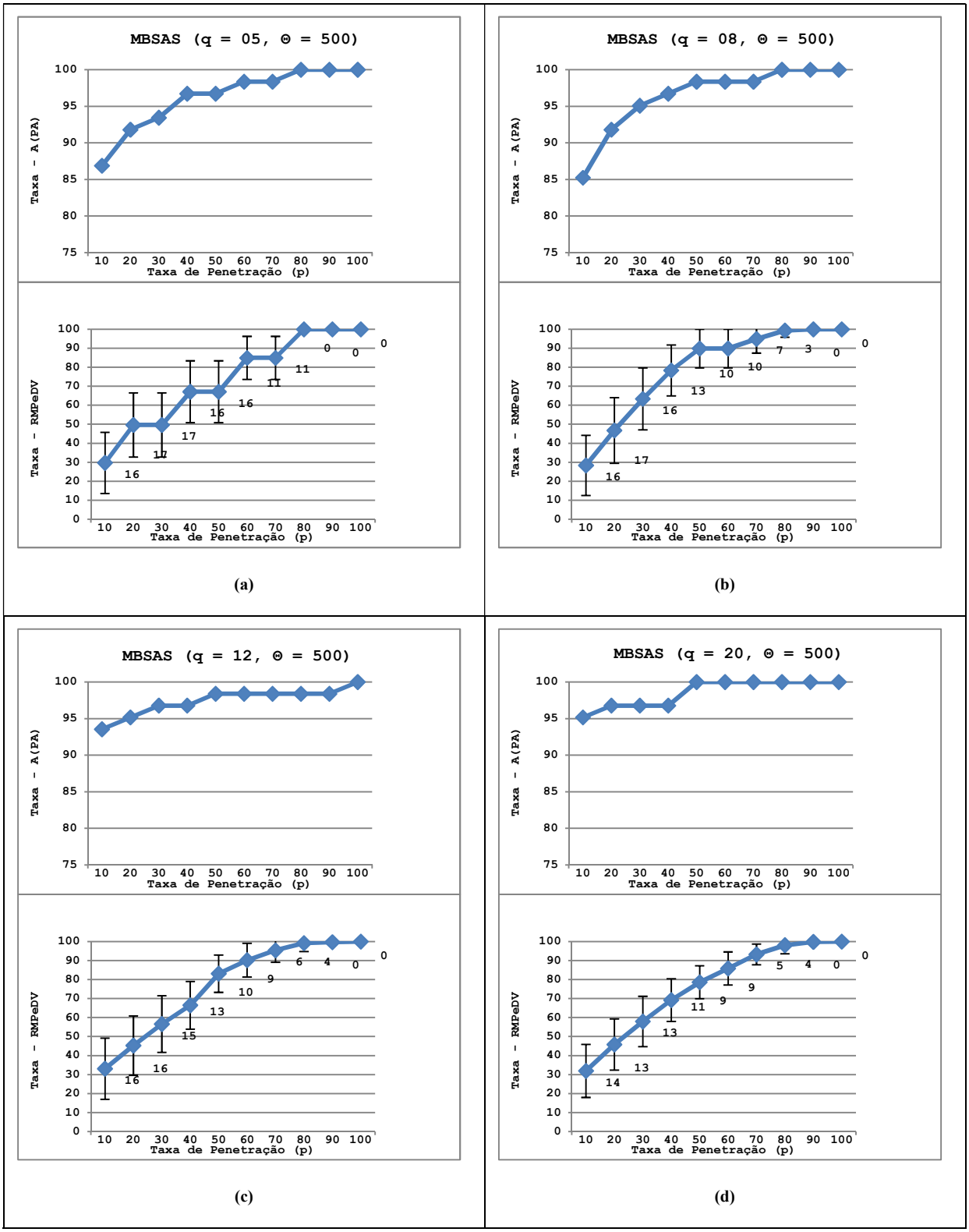

Figura 15. Resultados da busca reduzida pelo agrupamento do algoritmo MBSAS no parâmetro (Θ*)* com valor 500 e com variação de q (a-05, b-08, c-12 e d-20).

Tabela 8. Melhor resultado de cada agrupamento na menor taxa penetração no parâmetro (*Θ)* com valor 500 e com variação de q (05, 08, 12 e 20).

Resultado MBSAS				
Grupo/Limiar	**Percentual Taxa de penetração**	**Percentual Taxa de acerto**	**Percentual médio de Descritores percorridos na busca**	**Desvio padrão sobre o Percentual médio**
05/500	80	100	100	0
08/500	80	100	99,3	3,4
12/500	100	100	100	0
20/500	50	100	78,2	6,1

Nas figuras 15-b e 15-d tivemos os melhores resultados referenciados aos parâmetros utilizados no agrupamento da Figura 15. A taxa de penetração com maior redução foi alcançada na figura 15-d e com percentual médio de descritores percorridos em 78,2%, como demonstrado em tabela 8 (20/500). Para figura 15-b o percentual foi de 99,3%, diferença entre elas de apenas 21,1%. Figuras 15-a e 15-c apresentaram o pior desempenho, pois percorreram o banco por completo ou seja, da mesma forme que a busca exaustiva.

Na Figura 16, os resultados obtidos pelo algoritmo de busca por redução de espaço para descritores agrupados segundo o algoritmo MBSAS para variação do parâmetro de limiar de distância Θ com valor 1000.

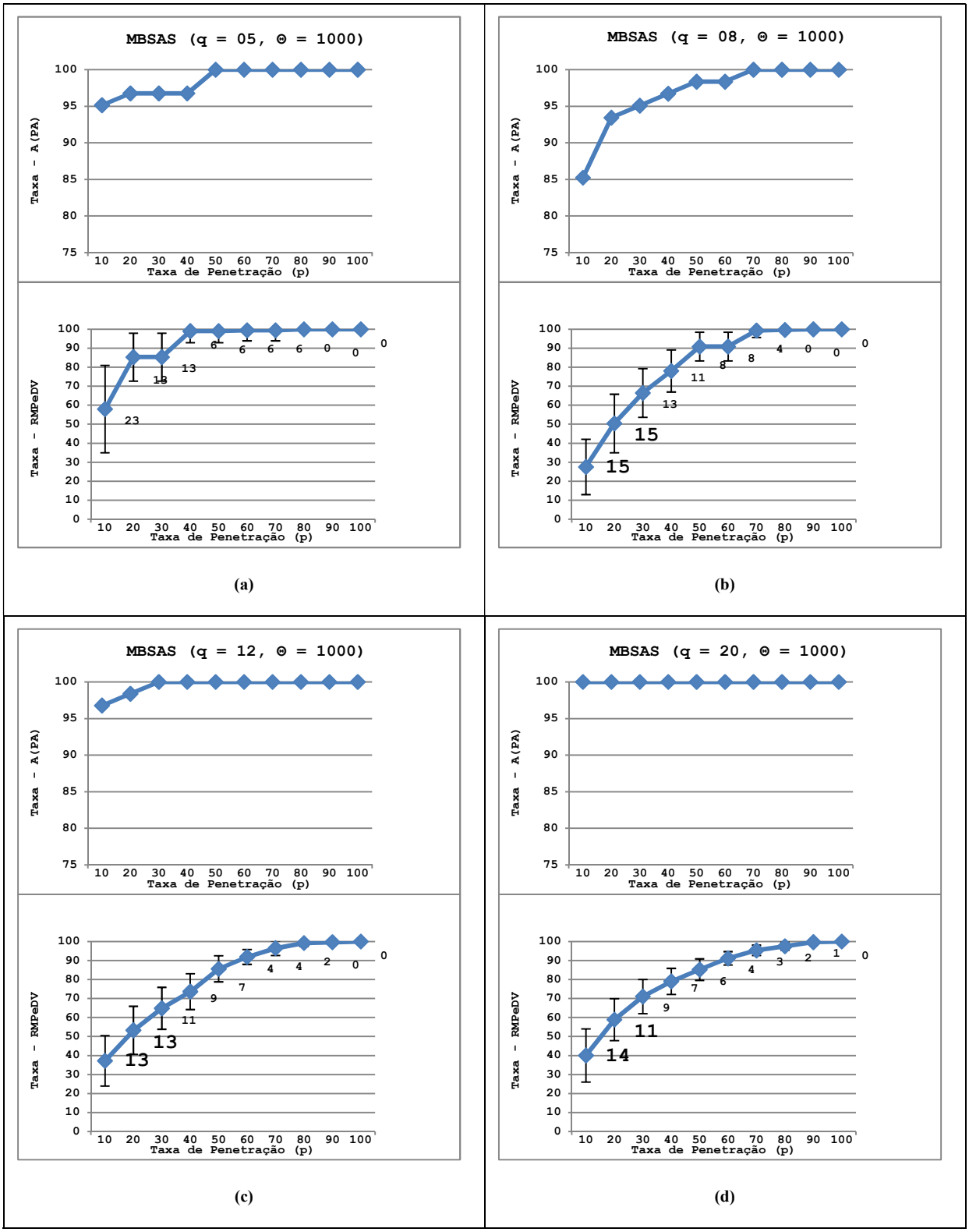

Figura 16. Resultados da busca reduzida pelo agrupamento do algoritmo MBSAS no parâmetro (Θ) com valor 1000 e com variação de q (a-05, b-08, c-12 e d-20).

Tabela 9. Melhor resultado de cada agrupamento na menor taxa penetração no parâmetro (*Θ)* com valor 1000 e com variação de q (05, 08, 12 e 20).

.Resultado MBSAS				
Grupo/Limiar	**Percentual Taxa de penetração**	**Percentual Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual Desvio padrão sobre o Percentual médio**
05/1000	50	100	99	6,1
08/1000	70	100	99,4	3,6
12/1000	30	100	64,8	11
20/1000	10	100	39,9	14

Os parâmetros utilizados na geração dos dados da Figura 16 apresentaram resultados satisfatórios. O melhor resultado foi atingido na figura 16-d, com a taxa de penetração com o menor nível, com apenas 10% da penetração no banco de dados e apenas 39,9% dos descritores percorridos. Na figura 16-c o resultado também foi satisfatório, com taxa de penetração igual a 30 e percentual médio de descritores percorridos igual a 64,8%. Para figuras 16-a e 16-b a taxa de penetração foi satisfatória porém o percentual médio de descritores percorridos não foi significativo como mostrado em tabela 9 ((05/100) e (08/1000)).

Tabela 10. Pior e melhor resultado para o agrupamento realizado pelo algoritmo MBSAS.

Resultado MBSAS					
Grupo/Limiar	**Percentual Taxa de penetração**	**Percentual Taxa de acerto**	**Percentual médio de Descritores percorridos na busca**	**Percentual Desvio padrão sobre o Percentual médio**	
12/500	100	100	100	0	Pior resultado
20/1000	10	100	39,9	14	Melhor resultado

A Tabela 10 nos mostra as duas situações de respostas do algoritmo de agrupamento MBSAS com o pior e o melhor resultado. Em tabela 10 (12/500) nenhum taxa de penetração menor que a busca exaustiva ou busca em 100% do banco de dados tiveram bons resultados. Enquanto que em tabela 10 (20/1000) o resultado foi excelente, em apenas 10% de penetração e 39,9% dos descritores percorridos foi atingido o mesmo

valor da busca exaustiva. Os parâmetros passados ao método de agrupamento MBSAS em tabela 10 (20/1000) geraram uma redução de 60,1% no tempo de busca.

6.6. Conclusão do melhor e pior resultado apresentado pelos algoritmos BSAS e MBSAS

A Tabela 11 apresenta o melhor e pior resultado obtido pelos testes realizados sobre os agrupamentos dos algoritmos de agrupamento sequencial BSAS e MBSAS.

Os dois algoritmos apresentaram praticamente o mesmo resultado final nos experimentos realizados neste trabalho, com apenas 0,1% de diferença entre eles. O melhor agrupamento se deu para o algoritmo MBSAS, com apenas 10% da taxa de penetração conseguiu atingir a mesma taxa de acerto da busca exaustiva e teve 39,9% dos descritores percorridos pela busca, resultado alcançado com q igual a 20 e limiar igual a 1000. O algoritmo BSAS para estas mesmas condições percorreu 40% dos descritos. E para o pior caso ambos tiveram o mesmo resultado.

Tabela 11. Comparativo entre o melhor e pior resultados de BSAS com os do MBSAS.

Resultado BSAS/MBSAS					
Algoritmo/ Grupo/ Limiar	**Percentual Taxa de penetração**	**Percentual Taxa de acerto**	**Percentual médio de Descritores percorridos na busca**	**Percentual Desvio padrão sobre o Percentual médio**	
BSAS 12/500	100	100	100	0	Pior resultado
BSAS 20/1000	10	100	40	14	Melhor resultado
MBSAS 12/500	100	100	100	0	Pior resultado
MBSAS 20/1000	10	100	39,9	14	Melhor resultado

Analisando que o algoritmo MBSAS é uma melhoria do BSAS e como seus resultados foram muito similares, BSAS tem a vantagem de agrupar em um único processamento e o MBSAS precisa processar duas vezes os dados, o que leva a consumir mais tempo de processamento. Porém, é necessário realizar testes com outros bancos de dados, podendo gerar variações nos resultados aqui obtidos. Estes testes serão aprofundados em trabalhos futuros com outros tipo de descritores e bancos maiores.

7. Conclusões

Este estudo apresentou um algoritmo de busca por redução de espaço (ABRE) em bancos de dados biométrico organizado em grupos pelos algoritmos de agrupamento sequencial BSAS e MBSAS. A proposta considera a redução do espaço de busca limitando a penetração nos agrupamentos, selecionando aqueles grupos mais similares ao descritor buscado. Foram apresentados os resultados de percentual de comparações de descritores para a busca assim como a precisão de acerto. Foram explorados valores diferentes dos parâmetros de agrupamentos para os algoritmos BSAS e MBSAS. Os resultados dos experimentos mostram um potencial de uso do algoritmo para reduzir na média mais de 50% o espaço de busca, mantendo a mesma taxa de precisão que no caso de uma busca exaustiva no bando de dados. Os resultados utilizando o algoritmo de agrupamento BSAS foram similares aos do MBSAS na taxa de penetração. Em alguns casos a variação se deu na taxa percentual média de descritores percorridos na consulta, porém, com valores bem próximos aos dois agrupamentos. Na comparação do pior e melhor caso, os dois algoritmos de agrupamento tiveram o mesmo resultado para precisão e no percentual médio de descritores percorridos a diferença foi 0,1% apenas. Este resultado mostra que o uso do algoritmo MBSAS apesar de ser uma melhoria do BSAS, em pequenos bancos de dados não traz vantagem significativa. O custo computacional é maior devido ao algoritmo realizar dois passos no processamento de todos os descritores no banco de dados. Trabalhos futuros devem ser realizados utilizando outros bancos de dados biométricos com números de descritores maiores e calculando as métricas repetidamente com a construção dos agrupamentos em ordem aleatória. Também será avaliado um terceiro algoritmo da mesma família, o TTSAS (*Two-Threshliaold Sequential Algorithmic Scheme*). Assim como explorar outros tipos de descritores, não apenas descritores faciais como os utilizados nesta pesquisa.

Referências

A. Jain and S. Pankanti, Advances in Fingerprint Technology, ch. Automated Fingerprint Identification and Imaging Systems. CRC Press, 2nd ed., 2001.

A. Mansfield and J. Wayman, "Best practices in testing and reporting performance of biometric devices," Aug 2002.

Bonato, C S, Neto, R M F, 2010. Um Breve Estudo Sobre Biometria. Departamento de Ciência da Computação, Universidade Federal de Goiás (UFG), Campus Catalão.

Bottou, L. and Bengio, Y. (1995). Convergence Properties of the KMeans Algorithm, Advances in Neural Information Processing Systems, 7, MIT Press, Denver, pp. 585-592.

Cappelli, R., Maio, D., Maltoni, D., Wayman, J. L., & Jain, A. K 2006. Performance Evaluation of Fingerprint Verification Systems. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Carpenter, G. A. and Grossberg, S. 1987. ART2: Self-organization of stable category recognition codes for analog input patterns, Applied Optics, (26), pp. 4919-4930.

Eastman, J. R.. Idrisi 2006. The Andes Edition. Worcester, MA: Clark University.

Das, N. 2003. *Hedge fund classification using K-means clustering method. In: 9th International Conference in Computing in Economics and Finance, University of Washington, Seattle. Disponível em:* < http://depts.washington.edu/sce2003/Papers/284.pdf >.

Florent Perronnin and Jean-Luc Dugelay 2005. Clustering Face Images with pplication to Image Retrieval in Large Databases, Institut Eur´ecom Multimedia ommunications Department 2229 route des Crêtes, BP 193 06904 Sophia-Antipolis C´edex, FRANCE.

F. Perronnin, J.-L. Dugelay, and K. Rose, "Deformable face mapping for person identification," in IEEE Int. Conf. on Image Processing (ICIP), 1, pp. 661–664, 2003.

Florent Perronnin and Jean-Luc Dugelay Clustering Face Images with Application to Image Retrieval in Large Databases Biometric Technology for Human Identification II, edited by Anil K. Jain, Nalini K. Ratha, Proc. of SPIE Vol. 5779 (SPIE, Bellingham, WA, 2005)

Gonzalez, R., Woods, Richard 2000. Processamento de imagens digitais, Editora Edgard Blucher LTDA.

Guan, B. X., Bhanu, B., Thakoor, N. S., Talbot, P. and Lin, S. 2013. Automatic Cell Region Detection By K-means With Weighted Entropy. In: IEEE 10th International Symposium on Biomedical Imaging: From Nano to Macro San Francisco, CA, USA, pp. 418-421.

Hansen, P., Jaumard, B 1997. Cluster Analysis and Mathematical Programming; Mathematical Programming no. 79. pp.191-215.

Hunny Mehrotra., Dakshina R., Kisku., Javdavpur University., V. Bhawani Radhika., Banshidhar Majhi., Phalguni Gupta 2010. Feature Level Clustering of Large Biometric Database

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

Jain, A., Hong, L., and Pankanti, S 2000. Biometric Identification, Comunications of the ACM, vol. 43, No. 2.

Jain, A., Ross, A. & Nandakumar, K. (2011), "Introduction to Biometrics", Springer US, 1ª Ed.

Jain, A.K., Murty, M.N. & Flynn, P.J 1999. "Data Clustering: A Review", ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

Jain A. K., Ross. A., Prabhakar. S 2004. An Introduction to Biometric Recognition, IEEE. Transaction on Circuits and Systems for Video Techonology Special Issue on Image and Video Based Biometrics v. 14, n, p. 4-20.

Jardini Evandro., Gonzaga Adilson., Traina Caetano Jr 2011. Metodologia para Indexação e Busca de Impressões Digitais através do uso de Função de Distância

Métrica. (IJACSA) International Journal of Advanced Computer Science and Applications.

Liu, C.L., Nakashima, K., Sako, H., Fujisawa H 2004. Handwritten digit recognition: Investigation of normalization and feature extraction techniques, Pattern Recognition, vol.37, no.2, pp. 265-279, Feb.

L. Hong and A. Jain, "Integrating faces and fingerprints for person identification," IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 20, pp. 1295–1307, Dec 1998.

Liu, M., Jiang, X. & Kot, A. C. (2007) "Efficient fingerprint search based on database clustering" , Pattern Recognition, 40, pp 1793 – 1803

Liu, S., e Silverman, M 2000. *A Practical Guide to Biometric Security Technology, IEEE Computer Society*, www.computer.org/itpro/homepage/Jan_Feb/security3.htm.

Maurya R., Singh S., Gupta P.R. and Sharma M. K. (2011). Road Extraction Using K-means Clustering and Morphological Operations. In: International Journal of Advanced Engineering Sciences and Technologies, v. 5, n. 2, pp. 290-295.

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

Mhatre A., Chikkerur S., and x Govindaraju V., 2005. "Indexing Biometric Databases Using Pyramid Technique", AVBPA05, pp. 841-849.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. Machine Learning, 39:103–134.

Novo, E. M. L. de M. 1992. Sensoriamento remoto: princípios e aplicações. São Paulo: Edgard Blucher.

Pedrycz, W., and Waletzky, J 1997. Fuzzy clustering with partial supervision. IEEE transactions on system, man and cybernetics, 27(5).

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

Poh, N., e Korczak, J 2001. Hybrid Biometric Person Authentication Using Face and Voice Features, Proceedings of the Third International Conference, Audio- and Video-based Biometric Person Authentication AVBPA 2001, Halmstad, Sweden.

Prabhakar, S., Pankanti, S., Jain, A. K 2003. Biometric Recognition: Security and Privacy Concerns. IEEE Security & Privacy. p. 33-42.

Putte, T., e Keuning, J 2000. Biometrical fingerprint recognition: don´t get your fingers burned, Proceedings of IFIP TC8/WG8.8 Forth Working Conference on Smart Card Research and Advanced Applications, Kluwer Academic Publishers.

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. M. E. (2014) Investigação de algoritmos sequênciais de agrupamento com pré-processamento de dados em aprendizado de máquina. Dissertação do PMCC-FACCAMP.

Reilo., Raul Sanches., Ávila., Carmem Sanches., Marcos., Ana Gonzalez 2000. Biometric Identification Trough Hand Geometry Measurements, IEEE Transactions on Pattern Analysis and Machine Intelligence,nº 10, vol. 22, p. 1168-1171.

Ribeiro I., Chiachia G., Marana N. A 2010. Reconhecimento de Faces Utilizando Análise de Componentes Principais e a Transformada Census, Universidade Estadual Paulista - UNESP (Campus de Bauru) Faculdade de Ciências - Departamento de Computação.

Rosa, R 2009. Introdução ao sensoriamento remoto. Uberlândia: EDUFU.

Samet, H. (2006) "Foundations of Multidimensional and Metric Data Structures", Morgan Kaufmann,

Sing, J.K. (2009) "A clustering and indexing technique suitable for biometric database" Master of Technology Thesis, Indian Institute of Technology, Kanpur.

Tatiraju S. and Mehta A. (2008). Image Segmentation using K-means clustering, EM and Normalized Cuts. In: Machine Learning Winter. Disponível em: <http://www.ics.uci.edu/~dramanan/teaching/ics273a_winter08/projects/avim_report.pdf>

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

Trahanias, P., Scordalakis, E 1989. An efficient sequential clustering method, Pattern Recognition, (22), n. 4, pp. 449-453.

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

Urtiga E. V. C., and Moreno E. D 2011. Keystroke-Based Biometric Authentication in Mobile Devices, , IEEE LATIN AMERICA TRANSACTIONS, VOL. 9, NO. 3.

Wessels, T., Omlin, C. W 2000. A Hybrid System for Signature Verification. IEEE. pp. 509-514.

Wirtz, B 1997. "Average Prototypes for stroke-based signature verification", ICDAR 97, Vol. 1, pp. 268 - 272, Germany.

Anexo I: Resultado Algoritmo BSAS

No anexo I, todas as evidencias geradas pelo agrupamernto realizado com BSAS.

Resultado BSAS				
Grupo/Limiar	**Percentual Taxa de penetração**	**Percentual Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
05/100	10	81,4	22,7	6,5
	20	89,8	43,4	9,4
	30	91,5	43,4	9,4
	40	94,9	64,5	9,3
	50	96,6	64,5	9,3
	60	98,3	83,4	5,9
	70	98,3	83,4	5,9
	80	100	99,8	2,0
	90	100	99,8	2,0
	100	100	100	0
08/100	10	79,7	19,2	8,2
	20	88,1	37,1	13,4
	30	91,5	52,9	14,4
	40	94,9	65,6	15,6
	50	96,6	77,2	12,7
	60	98,3	77,2	12,7
	70	98,3	85,4	8,5
	80	100	94,1	5,3
	90	100	99,7	2,0
	100	100	100	0
12/100	10	91,7	31,7	12,5
	20	95,0	45,1	13,7
	30	96,7	56,3	14,5
	40	98,3	67,5	12,4
	50	98,3	83,5	9,4
	60	100	90,3	9,1
	70	100	94,6	7,2
	80	100	98,8	3,9
	90	100	99,3	2,0
	100	100	100	0
20/100	10	90,2	26,4	10,0
	20	95,1	40,1	12,5
	30	96,7	52,0	12,7
	40	98,4	62,7	11,7
	50	98,4	73,0	9,8
	60	98,4	81,1	8,0
	70	98,4	88,9	6,6
	80	98,4	96,2	4,3
	90	100	99,1	1,5
	100	100	100	0

Resultado BSAS				
Grupo/Limiar	**Percentual Taxa de penetração**	**Percentual Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
05/300	10	81,4	22,7	6,5
	20	89,8	43,4	9,4
	30	91,5	43,4	9,4
	40	94,9	64,5	9,3
	50	96,6	64,5	9,3
	60	98,3	83,4	5,9
	70	98,3	83,4	5,9
	80	100	99,8	2,0
	90	100	99,8	2,0
	100	100	100,0	0
08/300	10	79,7	19,4	7,7
	20	86,4	38,1	11,5
	30	91,5	53,9	12,2
	40	94,9	68,8	10,8
	50	96,6	81,1	8,7
	60	98,3	81,1	8,7
	70	98,3	90,9	8,1
	80	100	95,6	5,0
	90	100	99,7	2,0
	100	100	100,0	0
12/300	10	91,7	30,1	12,0
	20	95,0	42,6	12,7
	30	96,7	55,2	11,6
	40	98,3	65,9	10,5
	50	98,3	83,0	8,9
	60	100	90,3	8,4
	70	100	95,0	6,4
	80	100	98,8	3,8
	90	100	99,3	2,0
	100	100	100,0	0
20/300	10	91,8	30,5	13,0
	20	96,7	44,3	13,6
	30	96,7	56,5	12,9
	40	98,4	67,3	11,8
	50	98,4	77,3	8,9
	60	98,4	85,0	8,8
	70	98,4	92,7	6,0
	80	100	97,9	4,5
	90	100	99,1	1,5
	100	100	100,0	0

Resultado BSAS				
Cluster/Limiar	**Taxa de penetração**	**Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
05/500	10	86,7	29,7	16,1
	20	91,7	49,6	16,9
	30	93,3	49,6	16,9
	40	96,7	67,0	16,3
	50	96,7	67,0	16,3
	60	98,3	84,8	11,4
	70	98,3	84,8	11,4
	80	100	99,8	2,0
	90	100	99,8	2,0
	100	100	100	0
08/500	10	85,0	28,4	15,8
	20	91,7	46,7	17,2
	30	95,0	63,3	16,3
	40	96,7	78,2	13,4
	50	98,3	89,7	10,2
	60	98,3	89,7	10,2
	70	98,3	94,7	7,4
	80	100	99,1	3,4
	90	100	99,7	2,0
	100	100	100	0
12/500	10	93,4	33,0	16,1
	20	95,1	45,1	15,6
	30	96,7	56,4	14,9
	40	96,7	66,3	12,6
	50	98,4	82,8	9,8
	60	98,4	89,9	8,9
	70	98,4	95,0	6,2
	80	98,4	98,8	4,4
	90	98,4	99,3	3,0
	100	100	100	0
20/500	10	95,1	31,8	14,0
	20	98,4	45,6	13,5
	30	98,4	57,7	13,2
	40	98,4	68,8	11,3
	50	100	78,6	8,6
	60	100	85,3	8,7
	70	100	92,6	5,4
	80	100	97,4	4,5
	90	100	99,0	0,7
	100	100	100	0

Resultado BSAS				
Cluster/Limiar	**Taxa de penetração**	**Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
05/1000	10	95,1	58,0	23,0
	20	98,4	85,3	12,6
	30	98,4	85,3	12,6
	40	98,4	98,9	6,1
	50	100	98,9	6,1
	60	100	99,3	5,5
	70	100	99,3	5,5
	80	100	99,8	2,0
	90	100	99,8	2,0
	100	100	100	0
08/1000	10	85,0	27,5	14,6
	20	93,3	50,3	15,4
	30	95,0	66,4	12,8
	40	96,7	77,9	11,1
	50	98,3	90,7	7,6
	60	98,3	90,7	7,6
	70	100	99,1	3,6
	80	100	99,4	3,4
	90	100	99,7	2,0
	100	100	100	0
12/1000	10	96,7	37,1	13,3
	20	98,4	53,2	12,6
	30	100	64,7	11,1
	40	100	73,4	9,4
	50	100	85,4	6,9
	60	100	91,6	3,9
	70	100	96,1	3,9
	80	100	98,8	1,8
	90	100	99,2	1,2
	100	100	100	0
20/1000	10	100	40	14,0
	20	100	58,7	11,1
	30	100	70,8	9,0
	40	100	78,7	6,8
	50	100	84,8	5,7
	60	100	90,6	3,5
	70	100	94,8	2,7
	80	100	96,8	2,0
	90	100	98,9	1,1
	100	100	100	0

Resultado BSAS				
Cluster/Limiar	**Taxa de penetração**	**Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
20/500	10	95,1	32,0	13,9
	20	100	46,0	13,4
	30	100	58,0	13,2
	40	100	69,0	11,3
	50	100	79,0	8,6
	60	100	86,0	8,7
	70	100	93,0	5,4
	80	100	98,0	4,4
	90	100	100	0
	100	100	100	0
20/1000	10	95,1	40,0	14,0
	20	100	59,0	11,0
	30	100	71,0	9,0
	40	100	79,0	6,8
	50	100	85,0	5,6
	60	100	91,0	3,5
	70	100	95,0	2,7
	80	100	97,0	2,0
	90	100	100	0
	100	100	100	0
20/1500	10	96,7	49,0	18,5
	20	96,7	65,0	11,5
	30	100	74,0	8,3
	40	100	82,0	6,0
	50	100	87,0	4,7
	60	100	92,0	3,6
	70	100	96,0	2,7
	80	100	99,0	1,8
	90	100	100	0
	100	100	100	0

Resultado BSAS				
Cluster/Limiar	**Taxa de penetração**	**Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
50/500	10	96,0	33,0	12,6
	20	100	51,0	12,4
	30	100	64,0	10,3
	40	100	78,0	8,1
	50	100	81,0	5,7
	60	100	87,0	4,7
	70	100	92,0	3,4
	80	100	96,0	1,9
	90	100	99,0	0,6
	100	100	100	0
50/1000	10	100	31,0	12,0
	20	100	47,0	12,4
	30	100	60,0	10,9
	40	100	71,0	7,9
	50	100	80,0	5,9
	60	100	86,0	3,7
	70	100	91,0	2,7
	80	100	95,0	1,8
	90	100	99,0	0,8
	100	100	100	0
50/1500	10	100	49,0	16,5
	20	100	65,0	12,6
	30	100	77,0	10,1
	40	100	84,0	9,8
	50	100	90,0	7,3
	60	100	94,0	5,5
	70	100	96,0	2,5
	80	100	97,0	1,8
	90	100	99,0	0,6
	100	100	100	0

Resultado BSAS				
Cluster/Limiar	**Taxa de penetração**	**Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
75/500	10	100	26,0	11,6
	20	100	44,0	12,8
	30	100	56,0	11,6
	40	100	67,0	11,3
	50	100	76,0	8,2
	60	100	83,0	5,6
	70	100	88,0	4,5
	80	100	94,0	2,7
	90	100	98,0	0,9
	100	100	100	0
75/1000	10	100	27,0	10,3
	20	100	46,0	12,0
	30	100	59,0	10,3
	40	100	71,0	6,5
	50	100	79,0	4,0
	60	100	86,0	2,7
	70	100	91,0	1,7
	80	100	95,0	1,0
	90	100	98,0	0,6
	100	100	100	0
75/1500	10	100	42,0	15,2
	20	100	62,0	11,4
	30	100	73,0	10,5
	40	100	82,0	9,4
	50	100	88,0	7,4
	60	100	93,0	3,7
	70	100	95,0	2,7
	80	100	97,0	1,9
	90	100	99,0	0,9
	100	100	100	0

Anexo II: Resultado Algoritmo MBSAS

No anexo I, todas as evidencias geradas peloo agrupamernto realizado com MBSAS.

Resultado MBSAS				
Cluster/Limiar	**Taxa de penetração**	**Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
05/100	10	81,7	22,8	6,5
	20	90,0	43,5	9,4
	30	91,7	43,5	9,4
	40	95,0	64,6	9,3
	50	96,7	64,6	9,3
	60	98,3	83,5	5,9
	70	98,3	83,5	5,9
	80	100	100	0
	90	100	100	0
	100	100	100	0
08/100	10	80,0	19,2	8,2
	20	88,3	37,2	13,4
	30	91,7	53,0	14,4
	40	95,0	65,8	15,6
	50	96,7	77,4	12,7
	60	98,3	77,4	12,7
	70	98,3	85,6	8,5
	80	100	94	5
	90	100	100	0
	100	100	100	0
12/100	10	93,4	31,7	12,5
	20	95,1	45,2	13,7
	30	96,7	56,5	14,5
	40	98,4	67,7	12,4
	50	98,4	83,7	9,4
	60	100	90	9,1
	70	100	94,9	7,2
	80	100	99,2	3,9
	90	100	100	2,0
	100	100	100	0
20/100	10	90,3	26,5	10,0
	20	95,2	40,3	12,5
	30	96,8	52,3	12,7
	40	98,4	63,0	11,7
	50	98,4	73,4	9,8
	60	98,4	81,6	8,0
	70	98,4	89,5	6,6
	80	98,4	96,9	4,3
	90	100	99,8	1,5
	100	100	100	0

Resultado MBSAS				
Cluster/Limiar	**Taxa de penetração**	**Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
05/300	10	81,7	22,8	6,5
	20	90,0	43,5	9,4
	30	91,7	43,5	9,4
	40	95,0	64,6	9,3
	50	96,7	64,6	9,3
	60	98,3	83,5	5,9
	70	98,3	83,5	5,9
	80	100	100	0
	90	100	100	0
	100	100	100	0
08/300	10	80,0	19,5	7,7
	20	86,7	38,2	11,5
	30	91,7	54,0	12,2
	40	95,0	69,0	10,8
	50	96,7	81,3	8,7
	60	98,3	81,3	8,7
	70	98,3	91,1	8,1
	80	100	95,9	5,0
	90	100	100	0
	100	100	100	0
12/300	10	91,7	31,7	12,5
	20	95,0	45,2	13,7
	30	96,7	56,5	14,5
	40	98,3	67,7	12,4
	50	98,3	83,7	9,4
	60	100	90,1	9,1
	70	100	94,9	7,2
	80	100	99,2	3,9
	90	100	99,8	2,0
	100	100	100	0
20/300	10	91,9	30,6	13,0
	20	96,8	44,5	13,6
	30	96,8	56,7	12,9
	40	98,4	67,7	11,8
	50	98,4	77,7	8,9
	60	98,4	85,5	8,8
	70	98,4	93,3	6,0
	80	100	98,6	4,5
	90	100	99,8	1,5
	100	100	100	0

Resultado MBSAS				
Cluster/Limiar	**Taxa de penetração**	**Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
05/500	10	86,9	29,7	16,1
	20	91,8	49,7	16,9
	30	93,4	49,7	16,9
	40	96,7	67,2	16,3
	50	96,7	67,2	16,3
	60	98,4	85,0	11,4
	70	98,4	85,0	11,4
	80	100	100	0
	90	100	100	0
	100	100	100	0
08/500	10	85,2	28,4	15,8
	20	91,8	46,8	17,2
	30	95,1	63,4	16,3
	40	96,7	78,4	13,4
	50	98,4	89,9	10,2
	60	98,4	89,9	10,2
	70	98,4	94,9	7,4
	80	100	99,3	3,4
	90	100	100	0
	100	100	100	0
12/500	10	93,5	33,0	16,1
	20	95,2	45,3	15,6
	30	96,8	56,6	14,9
	40	96,8	66,5	12,6
	50	98,4	83,1	9,8
	60	98,4	90,2	8,9
	70	98,4	95,3	6,2
	80	98,4	99,2	4,4
	90	98,4	99,7	3
	100	100	100	0
20/500	10	95,2	31,9	14,0
	20	96,8	45,8	13,5
	30	96,8	57,9	13,2
	40	96,8	69,2	11,3
	50	100	78,2	8,6
	60	100	85,8	8,7
	70	100	93,2	5,4
	80	100	98,1	4,5
	90	100	99,8	0,7
	100	100	100	0

Resultado MBSAS				
Cluster/Limiar	**Taxa de penetração**	**Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
05/1000	10	95,2	58,0	23,0
	20	96,8	85,4	12,6
	30	96,8	85,4	12,6
	40	96,8	99,0	6,1
	50	100	99,0	6,1
	60	100	99,5	5,5
	70	100	99,5	5,5
	80	100	100	0
	90	100	100	0
	100	100	100	0
08/1000	10	85,2	27,5	14,6
	20	93,4	50,4	15,4
	30	95,1	66,5	12,8
	40	96,7	78,1	11,1
	50	98,4	90,9	7,6
	60	98,4	90,9	7,6
	70	100	99,4	3,6
	80	100	99,7	3,4
	90	100	100	0
	100	100	100	0
12/1000	10	96,8	37,2	13,3
	20	98,4	53,3	12,6
	30	100	64,8	11,0
	40	100	73,6	9,4
	50	100	85,6	6,9
	60	100	91,9	3,9
	70	100	96,5	3,9
	80	100	99,2	1,8
	90	100	99,7	1,2
	100	100	100	0
20/1000	10	100	39,9	14,0
	20	100	58,9	11,1
	30	100	71,1	9,0
	40	100	79,0	6,8
	50	100	85,2	5,7
	60	100	91,2	3,5
	70	100	95,4	2,7
	80	100	97,5	2,0
	90	100	99,7	1,1
	100	100	100	0

Resultado MBSAS				
Cluster/Limiar	**Taxa de penetração**	**Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
20/500	10	97,0	31,6	13,9
	20	100	45,2	13,4
	30	100	57,0	13,2
	40	100	68,0	11,3
	50	100	77,1	8,6
	60	100	84,0	8,7
	70	100	91,1	5,4
	80	100	95,6	4,4
	90	100	97,1	0,7
	100	100	100	0
20/1000	10	98,0	39,7	14,0
	20	100	58,3	11,0
	30	100	70,2	9,0
	40	100	77,8	6,8
	50	100	83,7	5,7
	60	100	89,3	3,5
	70	100	93,3	2,7
	80	100	95,1	2,0
	90	100	97,0	1,1
	100	100	100	0
20/1500	10	100	48,2	18,5
	20	100	64,4	11,6
	30	100	73,5	8,3
	40	100	80,5	6,0
	50	100	86,0	4,7
	60	100	90,3	3,6
	70	100	94,1	2,7
	80	100	96,3	1,8
	90	100	97,1	0,6
	100	100	100	0

Resultado MBSAS				
Cluster/Limiar	**Taxa de penetração**	**Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
50/500	10	100	32,7	12,6
	20	100	50,2	12,4
	30	100	62,9	10,3
	40	100	72,9	7,0
	50	100	79,4	5,7
	60	100	84,8	4,7
	70	100	89,6	3,4
	80	100	93,9	1,9
	90	100	96,5	0,6
	100	100	100	0
50/1000	10	100	30,8	12,0
	20	100	46,6	12,4
	30	100	59,6	10,9
	40	100	70,1	7,9
	50	100	78,3	5,9
	60	100	84,7	3,7
	70	100	89,3	2,7
	80	100	92,8	1,8
	90	100	96,1	0,8
	100	100	100	0
50/1500	10	100	48,2	16,4
	20	100	64,7	12,6
	30	100	75,8	10,1
	40	100	83,3	9,8
	50	100	88,8	7,3
	60	100	91,9	5,5
	70	100	93,9	2,5
	80	100	95,0	1,7
	90	100	96,4	0,6
	100	100	100	0

Resultado MBSAS				
Cluster/Limiar	**Taxa de penetração**	**Taxa de acerto**	**Percentual de Descritores percorridos na busca**	**Percentual desvio padrão**
75/500	10	97,0	26,2	11,6
	20	100	43,6	12,8
	30	100	54,8	11,6
	40	100	66,2	11,3
	50	100	74,5	8,2
	60	100	81,6	5,6
	70	100	86,2	4,5
	80	100	91,3	2,7
	90	100	95,4	0,9
	100	100	100	0
75/1000	10	100	26,4	10,3
	20	100	45,4	12,0
	30	100	58,6	10,3
	40	100	70,3	6,5
	50	100	78,0	4,0
	60	100	84,3	2,7
	70	100	88,6	1,7
	80	100	92,9	1,0
	90	100	95,4	0,6
	100	100	100	0
75/1500	10	100	42,0	15,0
	20	100	61,1	11,4
	30	100	72,1	10,4
	40	100	80,7	9,3
	50	100	86,5	7,4
	60	100	91,0	3,7
	70	100	93,3	2,7
	80	100	94,7	1,9
	90	100	96,2	0,9
	100	100	100	0