Uma técnica de agrupamento de elementos baseada em *Tolerance Near Sets*

Diego Saqui, José H. Saito*
Departamento de Computação - DC
Universidade Federal de São Carlos - UFSCar
São Carlos, SP, Brasil
(*) FACCAMP - Campo Limpo Paulista, SP, Brasil
Email: diego.saqui@dc.ufscar.br,saito@dc.ufscar.br

Lúcio A. de C. Jorge, Rodrigo B. Piassi
Embrapa Instrumentação Agropecuária
São Carlos, SP, Brasil
Email: lucio.jorge@embrapa.br,rodrigopiassi@outlook.com

***Abstract*—This study aims to explore the Near Sets and Tolerance Near Sets methodologies for clustering task. The groups generated by the approach should be used to represent classes and allow the application of classification methods. This work has been developed and tested for mango color classification and significant results were obtained. One problem noted was the large number of groups generated and there is need for adaptations of the method tested previously. The method is currently under development and is based on methodology of Tolerance Near Sets and statistical approaches. It is expected that the approach to be developed presenting significant results in the clustering process.**

***Abstract*—Este trabalho tem o objetivo de explorar a metodologia de *Near Sets* e *Tolerance Near Sets* na tarefa de agrupamento de elementos. Os grupos gerados pela abordagem, devem ser utilizados para representação de classes e permitir a aplicação de métodos de classificação. Este trabalho foi desenvolvido e testado para o agrupamento de cores de mangas, onde foram obtidos resultados significativos. Um problema observado, foi o grande número de grupos gerados e existe necessidade de adaptações do método de agrupamento previamente testado. O método atualmente está em desenvolvimento e é baseado na metodologia de *Tolerance Near Sets* e abordagens estatísticas. Espera-se que a abordagem a ser desenvolvida apresente resultados significativos no processo de agrupamento.**

***Keywords*-Near Sets; Tolerance Near Sets; Clustering.**

I. Introdução

Nas áreas de visão computacional e reconhecimento de padrões, técnicas de agrupamento (clustering) têm sido utilizadas na geração de grupos de forma não supervisionada e aplicadas em diferentes estudos, tais como seleção automática de frutas [1], explosões solares [2], entre outras. Cada agrupamento pode ser utilizado para representar elementos com diferentes características, como por exemplo, tipos de frutas. Técnicas de agrupamento são categorizadas como abordagens de aprendizado de máquina não supervisionado e são utilizadas para criar grupos por meio do estabelecimento de relacionamentos entre elementos. Técnicas de agrupamento buscam encontrar uma estrutura em suas entradas, não contendo informações prévias sobre os grupos a serem formados [3]. Dentre essas técnicas são comuns o *single/complete-linkage clustering* [4], o *K-means* [5], e mais recentemente abordagens baseadas em *Near Sets* [6].

A abordagem explorada neste trabalho, é baseada em *Near Sets* e *Tolerance Near Sets (TNS)* para o agrupamento de elementos. *Near Sets* provê formalidades para identificar, comparar e mensurar a semelhança de objetos (elementos) com base em suas características. *TNS*, uma sub-categoria de *Near Sets*, utiliza um limiar de tolerância para estabelecer relações entre os objetos [6]. Em relação a outras abordagens como *K-means*, métodos baseados em *TNS* tem se apresentado de forma eficiente no processo de agrupamento [2].

Near Sets quando aplicado na comparação de imagens opera sobre seus descritores determinando a diferença entre as imagens em comparação. Essa diferença é mensurada (por equações como distância euclidiana), e o resultado obtido é verificado por um limiar de tolerância, indicando se os elementos são semelhantes. Uma das vantagens dessa técnica é a possibilidade de comparar vários objetos entre si em um determinado contexto.

Neste trabalho, foi desenvolvido uma metodologia baseada em *TNS* para geração de grupos de classes utilizando atributos extraídos de imagens visando a segmentação. O algoritmo com a metodologia proposta foi testado em imagens de mangas visando a segmentação por qualidade de frutas baseado em cores.

Contribuições: Este trabalho propõe uma abordagem, baseada em *TNS* para tarefa de agrupamento de elementos (imagens), se preocupando com o problema de geração excessiva de grupos. A tarefa de agrupamento já é realizada e a parte do método que está em desenvolvimento é a redução do número de grupos por meio de junção.

A. Trabalhos Relacionados

Algoritmos de agrupamento são frequentemente explorados na área de aprendizado de máquina e podem ser categorizados como hierárquicos como *single linkage clustering* [4], baseados em particionamento como o *K-means* [5], baseados em modelos como *Expectation-Maximization* [7], entre outros.

Em algoritmos hierárquicos como o *single-linkage clustering* e o *complete-linkage clustering*, o agrupamento parte do princípio onde inicialmente existe um elemento por classe. A partir desse princípio são estabelecidas as semelhanças entre as classes através dos elementos contidos em cada classe inicial.

O processo de comparação entre classes é o que diferencia as técnicas de *single-linkage clustering* e o *complete-linkage clustering*. O *single-linkage clustering* considera a distância entre os menores elementos de cada classe enquanto que o *complete-linkage clustering* as maiores distâncias. Ainda nessa categoria existe o *average-linkage clustering* que considera a distância entre classes através da média de alguns elementos [4].

O *K-means* é um algoritmo clássico de mineração de dados que permite estabelecer K grupos em um processo de agrupamento. Esse algoritmo consiste em particionar uma população de N elementos (por meio de suas características) em K conjuntos. Elementos com características semelhantes são inseridos no mesmo conjunto [5].

Baseado em *Near Sets*, o método *Tolerance Near Sets (TNS)* tem sido explorado na literatura para o desenvolvimento de algoritmos de agrupamento de imagens, como no contexto de explosões solares [2]. Essa abordagem consiste na busca de combinações de grupos possíveis para diferentes objetos e considera a possibilidade de um objeto estar contido em mais de uma classe. O método desenvolvido foi comparado com o *K-means* e apresentou resultados significativos em relação a essa técnica. A aplicação de *TNS* demonstrou características importantes em processos de comparação, como a possibilidade de um objeto estar contido em mais de uma classe e a utilização do limiar de tolerância, e por isso é um interessante campo de pesquisa.

II. Fundamentos técnicos

A. Near Sets

Near Sets é um fundamento teórico que descreve que dois objetos são similares se eles contêm características com as mesmas descrições [2]. *Near Sets* funciona como um *framework* matemático e opera como a percepção humana. *Near Sets* considera que objetos são similares se tiverem características comuns. Objetos de conjuntos em *Near Sets* são identificados usando uma relação de indiscernibilidade [6]. Algumas definições de *Near Sets* úteis para este trabalho são apresentadas a seguir:

Definição 1. Objeto: Um objeto x é algo que existe no mundo real que pode ser percebido e descrito por suas características, por exemplo, imagens obtidas por câmeras que podem ser discriminadas por descritores.

Definição 2. Função de descrição: A função de descrição é responsável por retornar um valor que representa uma característica perceptível do objeto. O conjunto dessas funções é responsável pela descrição do objeto.

Definição 3. Sistema perceptivo: Um sistema perceptivo $< O, F >$ consiste de um conjunto não vazio O de objetos de amostra e um conjunto não vazio F de funções reais/descrição, $\phi \in F$ tal que $\phi : O \rightarrow \mathbb{R}$. Sistemas perceptivos são aplicados na comparação de novos objetos.

Definição 4. Descrição do objeto: A descrição do objeto é um vetor composto por um conjunto de funções de descrição B, sendo que $B \subset F$. O vetor de descrição de um objeto perceptual $x \in O$ é apresentado na equação (1)

$$\phi B(x) = (\phi 1(x), \phi 2(x), ...\phi i(x), ..., \phi l(x)) \quad (1)$$

sendo l o tamanho do vetor ϕB, e cada $\phi i(x)$ em $\phi B(x)$ uma função de descrição que é parte da descrição do objeto $x \in O$. Esse vetor é útil para descrever objetos por suas características.

Definição 5. Relação de indiscernibilidade e relação de indiscernibilidade fraca: A relação de indiscernibilidade perceptual determina a diferença entre as características dos objetos que estão sendo comparados. A relação de indiscernibilidade perceptual B é representada na equação (2).

$$B(x) = (x, y) \in O \times O : \forall \phi i \in B.\phi i(x) = \phi j(x) \quad (2)$$

Essa relação permite determinar a equivalência entre objetos e permitindo agrupar os que possuem características semelhantes. Considerando $\phi i \exists F$ é importante determinar a relação de indiscernibilidade perceptual fraca ϕi, que é representada na equação (3).

$$\phi i(x) = (x, y) \in O \times O : \exists \phi i \in B.\phi i(x) = \phi j(x) \quad (3)$$

Essa relação, diferentemente da relação de indiscernibilidade perceptual que considera a similaridade entre objetos com base em todas características, indica que a similaridade pode ser indicada por apenas algumas características.

Definição 6. *Tolerance Near Sets (TNS)*: TNS fornece uma base matemática para a definição de similaridade entre um par de objetos. Essa definição é dada por um espaço de tolerância que é um grau de relaxamento para uma relação de indicernibilidade. Um espaço de tolerância $< X, \xi >$ é constituído de uma relação de tolerância ξ sobre $X \subseteq O$ que é reflexiva e simétrica. Sendo $\xi \in \mathbb{R}$, para cada $B \subseteq F$ a relação de tolerância perceptual é definida conforme a equação (4)

$$\cong B, \varepsilon = (x, y) \exists O \times O : ||\phi(x) - \phi(y)||_2 < \varepsilon. \quad (4)$$

Essa relação em conjunto com as outras definições são utilizadas na metodologia proposta, onde a aplicação e os resultados são apresentados nas seções seguintes.

III. Metodologia e Experimentos

A metodologia utilizada neste trabalho segue os princípios de *Tolerance Near Sets* considerando descritores de cores (luminosidade e índices de dimensões de cores) como características a serem comparadas e uma etapa de agrupamento de imagens diferenciada.

Com um conjunto de 200 imagens de mangas em diferentes estados de maturação foram estabelecidos dois sub-conjuntos, um para amostras e outro para validar a abordagem proposta. Dessas imagens, 100 foram utilizadas para através de um processo de particionamento, gerar 1000 sub-imagens de amostras de tamanho 30 $\times$ 30 pixels e as outras 100 imagens para validação do sistema. Cada uma das sub-imagens de amostra foi considerada como um objeto no contexto de *Near Sets*.

A função de descrição utilizada foi a média de cores dos pixels de cada característica do espaço de cores Lab de cada

sub-imagem, ou seja, média de L, média de a e média de b. Para estabelecer a relação de indiscernibilidade perceptual, no processo de comparação foi utilizado o cálculo de distância euclidiana entre as imagens (objetos) comparadas.

Por final foi estabelecida uma relação de tolerância com um limiar que indica se duas sub-imagens pertencem ao mesmo grupo ou não. Essa escolha aconteceu de forma empírica com alguns testes por especialistas através de diferentes limiares. Após essas definições, são apresentadas as fases que consistiram deste trabalho.

Fase 1 - Pré-processamento: As imagens foram separadas em dois conjuntos, sendo um para gerar dados de amostra e outro de testes. O processo de geração das sub-imagens foi executado nesta etapa, onde as sub-imagens que representavam o fundo foram descartadas. Com as sub-imagens geradas, um método de extração de descritores Lab e formulação da média para cada descritor (funções de descrição em *Near Sets*), foram aplicados e as informações armazenadas.

Fase 2 - Agrupamento de objetos: Nessa fase um processo de agrupamento foi realizado onde cada sub-imagem é comparada com todas outras utilizando a média dos atributos Lab (funções de descrição) e com o uso de distância euclidiana (relação de indiscernibilidade). De acordo com o limiar de tolerância é indicado se duas sub-imagens que estão sendo comparadas são de mesma classe ou não. Os grupos gerados, podem posteriormente ser nomeados e utilizados como classes de sub-imagens. Com base na aplicação dessa metodologia algumas situações são representadas na Figura 1.

Na Figura 1 as sub-imagens (30x30 pixels) com números representam objetos e as elipses representam grupos/classes. Nesse cenário cada imagem é comparada com todas as outras por meio de seus descritores e distância euclidiana. O resultado da distância euclidiana para cada comparação é confrontado com um limiar de tolerância para estabelecer se dois objetos são de mesmo grupo ou não. A Figura 1a é baseada em um agrupamento de imagens de mangas com *Tolerance Near Sets*, onde foi utilizado um limiar com valor 0.11, e a Figura 1b representa o mesmo contexto porém com um limiar 0.15. Nas Figuras 1a e 1b é possível notar que o valor do limiar em *Tolerance Near Sets* é responsável por controlar a pertinência de um determinado objeto em uma classe. Com um limiar maior como na Figura 1b as classes abrangem um número maior de objetos (imagens 1, 2, 3 e 5 em um mesmo grupo), enquanto com um limiar menor, como na Figura 1a, as classes são mais restritivas em relação aos objetos (imagens 1 e 3 em um grupo, 2 e 3 em outro e a 5 em nenhum grupo).

Fase 3 - Classificação: Com o intuito de analisar os grupos de classes gerados, essa fase consiste em classificar novas mangas. Novas mangas são particionadas (sub-imagens de 30 $\times$ 30 pixels) e são extraídos seus descritores (média de cores Lab) assim como na etapa de geração de amostras. Cada sub-imagem é comparada com os valores médios dos descritores das amostra de cada uma das classes existentes. Essa comparação é realizada com a equação de distância euclidiana para cada classe, e a classe que rotula a nova sub-imagem é que apresenta a menor valor de distância.

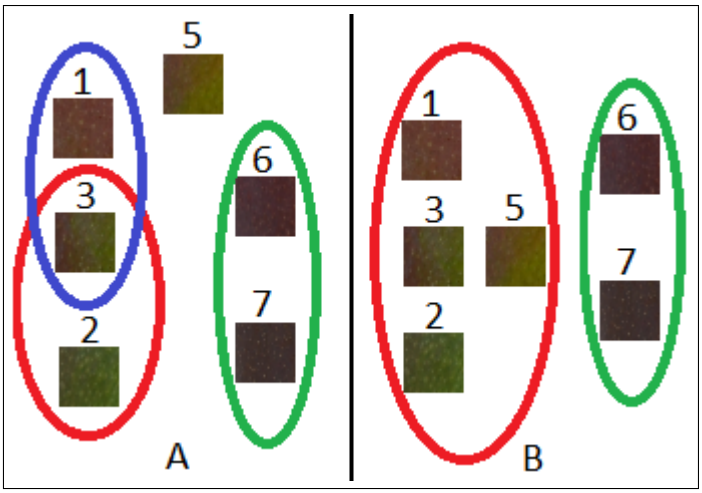

Fig. 1. A. Conjuntos/Classes geradas para alguns dados de amostra com um limiar de 0.11. B. Conjuntos/Classes geradas para alguns dados de amostra com um limiar de 0.15.

IV. Resultados iniciais e discussão

Para validação da abordagem, foi observada a relação do limiar de tolerância (obtido empiricamente) e as classes geradas. Para um limiar de 0.009 foram geradas 235 classes e para um limiar de 0.11 foram geradas 1513 classes. Observou-se que o limiar de tolerância menor (0.009), fez com que o agrupamento fosse mais restrito, gerando classes com um número menor de sub-imagens. A vantagem da utilização de um limiar menor é maior velocidade no processo e que devido ao número de classes geradas ser menor, posteriormente, processos de classificação podem ser executados com maior velocidade. A desvantagem de um limiar menor é que os dados ficam muito ajustados ao conjunto de treinamento (*overfitting*), e na classificação podem não abranger todos os casos. Para ambos os limiares é notável um número alto de classes geradas.

Outra observação foi realizada em um processo prático de classificação, considerando as classes geradas pela abordagem de *Tolerance Near Sets* e por análise de valores falso-positivos e verdadeiro-positivos obtidos. Considerando o limiar de tolerância de 0.11 (por não sofrer *overfitting*), para uma classificação de 116 sub-imagens de mangas foram obtidos 95 valores verdadeiro-positivos (classificações corretas) e 21 valores falso-positivos (classificações de forma errada). Com o conjunto de dados corretamente (verdadeiro-positivos) e erroneamente (falso-positivos) classificados, o resultado da avaliação nesse critério é apresentado no gráfico de Característica de Operação do Receptor (do inglês *Receiver Operating Characteristic - ROC*) apresentado na Figura 2.

No gráfico da Figura 2 é possível observar que a taxa de valores verdadeiro-positivo foi maior que 0.75 e a taxa de falso-positivos foi que menor que 0.1 que podem ser considerados bons resultados no processo de classificação. Com uma observação empírica foi notado que os 21 valores falso-positivos normalmente estavam relacionados com sub-imagens de cantos de mangas e provavelmente utilizando mais

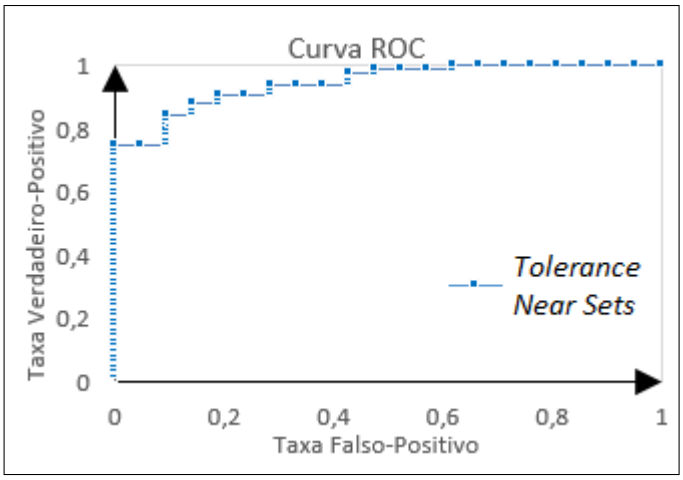

Fig. 2. Curva ROC para classificação de mangas com a abordagem proposta utilizando um limiar de tolerância de 0.11

descritores (como textura) no processo de geração de classes esse número seria reduzido.

V. Problemas e Adaptações Futuras

Problemas: Devido ao grande número de grupos gerados, pretende-se desenvolver um método de junção de grupos ou adaptar o método considerado para limitar o número de grupos gerados. Outros métodos como o *single-linkage clustering* e o *complete-linkage clustering* foram análisados, porém demonstraram o problema de considerarem situações extremas na comparação de imagens. Considerando que em um grupo existe uma sequência de imagens semelhantes com tonalidades em azul e uma única imagem verde, e em outro grupo existe uma sequência de imagens semelhantes com tonalidades em amarelo e uma única imagem verde, se as imagens verdes de cada classe tiverem a menor ou maior valor de distância (euclidiana ou outra) o *single-linkage* ou o *complete-linkage* irão agrupar esses grupos erroneamente.

Adaptação: A proposta de desenvolvimento desse método de agrupamento, considera os conceitos de *Tolerance Near Sets*. Nessa abordagem será considerado a média de cada descritor de imagem de cada grupo, e um limiar, aplicado a cada descritor, determinado por uma margem baseada no erro amostral de cada grupo. Para o caso das mangas com descritores Lab a equação (5) representa o processo de obtenção da média e erro amostral de um determinado grupo.

$$\bar{L} \pm z_l \cdot \frac{\sigma_l}{\sqrt{(n)}}, \bar{a} \pm z_a \cdot \frac{\sigma_a}{\sqrt{(n)}}, \bar{b} \pm z_b \cdot \frac{\sigma_b}{\sqrt{(n)}} \qquad (5)$$

$\bar{L}$, $\bar{a}$, $\bar{b}$ são as médias, z_l, z_a, , z_b os níveis de confiança, σ_l, σ_a, σ_b os desvios padrão respectivamente de cada um dos parâmetros Lab e n o número de objetos para um determinado grupo. Dessa forma, por exemplo, se a média Lab de um grupo $C1$ estiver dentro do intervalo (média e erro amostral) de um grupo $C2$, eles serão agrupadas. No algoritmo a ser desenvolvido, pretende-se que cada grupo seja comparado com todos os outros na busca do que possui menor variação e que um grupo já agrupado não seja observado novamente. Espera-se que este algoritmo seja capaz de reduzir o número de grupos a cada execução e também que apresente melhores resultados em relação aos já mencionados.

VI. Conclusão

Este artigo explora e propõe a utilização de uma metodologia baseada em *Tolerance Near Sets* para o agrupamento de imagens. Para validar a metodologia foi estabelecido um sistema com o propósito de geração de grupos de cores de mangas, e com os grupos gerados, foi considerado a classificação de novas imagens. Durante os testes foi realizado uma análise dos grupos gerados e as classificações de novas imagens com base nesses grupos. Em 116 classificação foram obtidas 95 corretas contra 21 classificações erradas. Dentre as imagens classificadas corretamente, o sistema foi capaz de classificar imagens mescladas (parcialmente verde e parcialmente vermelha) que é uma característica importante. Desta forma, os resultados apresentados foram considerados significativos, permitindo ainda melhorias com a utilização de novos descritores. Para melhorar o desempenho do sistema na questão de número de grupos gerados, que causa um alto custo computacional no processo de classificação, pretende-se desenvolver ou ajustar o método considerado para junção ou limitação do número de grupos. Esse processo deverá reduzir o número de grupos do sistema de forma adequada, permitindo melhoria no processo de classificação.

Agradecimentos

Agradecemos o fornecimento de recursos do Laboratório de Imagens da EMBRAPA Instrumentação e o CNPq com os projetos de n^{o}s 310310/2013-0 e 403426/2013-8 por proporcionarem recursos para o desenvolvimento desta pesquisa.

References

[1] H. Zawbaa, M. Abbass, M. Hazman, and A. Hassenian, "Automatic fruit image recognition system based on shape and color features," in *Advanced Machine Learning Technologies and Applications*, ser. Communications in Computer and Information Science, A. Hassanien, M. Tolba, and A. Taher Azar, Eds. Springer International Publishing, 2014, vol. 488, pp. 278–290. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-13461-1_27

[2] G. Poli, E. Llapa, J. Cecatto, J. Saito, J. Peters, S. Ramanna, and M. Nicoletti, "Solar flare detection system based on tolerance near sets in a gpucuda framework," *Knowledge-Based Systems*, vol. 70, no. 0, pp. 345 – 360, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095070511400269X

[3] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

[4] S. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967. [Online]. Available: http://dx.doi.org/10.1007/BF02289588

[5] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[6] J. F. Peters and P. Wasilewski, "Foundations of near sets," *Information Sciences*, vol. 179, no. 18, pp. 3091 – 3109, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025509001960

[7] J. Han, M. Kamber, and J. Pei, *Data Mining, Southeast Asia Edition: Concepts and Techniques*, ser. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2006. [Online]. Available: https://books.google.com.br/books?id=AfL0t-YzOrEC