MAC – 5701 Tópicos em Ciência da Computação

Prof. Responsável: Yoshiharu Kohayakawa

Relatório de Estudos

Aluno: **Germano Capistrano Bezerra**

Prof. Orientador: **Marcelo Finger**

Tema: Aprendizado Computacional Não-Supervisionado e Métodos de *Clustering*

Introdução

No contexto de sistemas distribuídos, surge o conceito de computação em grade, que consiste no uso simultâneo de recursos de vários computadores em uma rede para a resolução de um problema. Grades de computadores podem ser usadas para otimizar a utilização de recursos computacionais ociosos, distribuídos ao longo de diversas redes de computadores. Nestes casos, a idéia básica é permitir que máquinas sejam conectadas à grade, compartilhando e disponibilizando seus recursos, socializando o seu uso pelos diversos processos a serem executados na rede.

O projeto de pesquisa InteGrade, em desenvolvimento no IME-USP, visa o estudo e implementação de uma infra-estrutura para grades de computadores. Há diversas linhas de pesquisa sendo abordadas, a saber:

- requisitos e arquitetura geral do sistema;
- paralelização de problemas computacionalmente difíceis;
- segurança;

- identificação de padrões de acesso de usuários e de disponibilidade de recursos; e
- mobilidade de código e computação ubíqua.

Particularmente, a identificação de padrões de uso dos recursos disponíveis é um processo de aprendizado constante. Esse aprendizado tenta estabelecer métricas que permitam à grade a identificação da disponibilidade dos recursos compartilhados, possibilitando uma melhor distribuição de carga entre os nós envolvidos no sistema.

Para tanto, é necessário monitoramento de alguns dos recursos das máquinas envolvidas na grade, como:

- memória física disponível;
- memória disponível na área de *swap*;
- quantidade de espaço livre em disco; e
- percentual de utilização do processador.

Com essas informações, deverá ser possível o estabelecimento de categorias de comportamento das máquinas. Em um dado momento de solicitação de recursos de uma dada máquina, é identificada qual a categoria em operação corrente, permitindo uma decisão sistemática sobre a possibilidade de utilização do recurso compartilhado.

Objetivo do Estudo

Este estudo tem por objetivo a fundamentação teórica e a realização de um levantamento de métodos de aprendizado computacional visando a implementação da identificação de padrões de acesso de usuários e de disponibilidade de recursos, no escopo do projeto de pesquisa InteGrade.

Métodos de Aprendizagem Computacional

Aprendizagem Computacional é genericamente referida como o estudo de mecanismos utilizados por sistemas inteligentes visando melhorar seus desempenhos com o tempo. Esses mecanismos estão intimamente relacionados com a aquisição de conhecimento a partir da experiência em algum ambiente.
Pat Langley [Lan96] sugere um arcabouço para a implementação de aprendizagem computacional baseado nesta definição, especificamente em quatro termos básicos:

- Desempenho – medida quantitativa de algum aspecto relacionado com o domínio da informação em que o estudo é desenvolvido. A melhoria de desempenho está relacionada com mudanças capturadas por essas medidas;
- Ambiente – contexto em que se define o problema a ser analisado;
- Conhecimento – organização dos dados em alguma estrutura interna obtida da experiência através da manipulação de dados por sistema ou intervenção manual;
- Aprendizado – objetivo a ser alcançado com a organização dos outros componentes.

O Ambiente

A contextualização do ambiente em que ocorre o aprendizado depende de alguns fatores. O primeiro deles diz respeito à definição do objetivo de melhoria de desempenho a ser atingido. Todo aprendizado é feito com a intenção de maximizar ou minimizar alguma medida que faça sentido ao tema de estudo.

Outro fator é o modo (ausência ou presença) de supervisão do processo. Em trabalhos de classificação de dados, o ambiente supervisionado tem disponível um conjunto de dados de treinamento já pré-classificados. Na modalidade de treinamento que lida com a resolução de problemas, esse ambiente conta com o auxílio de um oráculo que determina os resultados esperados. No ambiente não-supervisionado não há presença de tipo algum de tutor.

Também faz parte do ambiente a forma de aquisição dos dados. Ela pode ser pontual, com a obtenção de todas as informações simultaneamente, antes do início do procedimento de análise dos dados, ou constante, com a obtenção de uma quantidade de informação por vez. Há também a possibilidade de um modo de operação híbrido: é feita uma coleta inicial de uma boa quantidade de dados, sendo processados logo em seguida, e, com o decorrer do tempo, novos eventos geram mais dados que são coletados num procedimento constante. O ambiente da InteGrade prevê um processo de aprendizado constante, mas assume-se que um conjunto pré-definido de dados já foi coletado para cada nó da Grade.

O último fator de ambiente diz respeito à regularidade do mesmo: relevância das informações obtidas, presença de ruídos nos dados de entrada e a consistência do ambiente ao longo do tempo.

Natureza da Representação da Experiência

A representação da experiência é fundamental para que o processo de aprendizado possa ser capaz de construir conhecimento a partir dela. A experiência, aqui definida como o conjunto de observações realizadas em algum domínio do conhecimento, normalmente diz respeito à medição de alguma grandeza:

- binária – que mede a presença ou ausência de alguma característica;
- atributos enumerados – similar às grandezas binárias, mas que permite mais de dois valores mutuamente exclusivos;
- numérica – valor inteiro ou real que representa alguma medida realizada. Essa é a forma de representação mais simples de ser usada. Normalmente, cada objeto analisado (ou medido) é composto de um número n de características. Assim, a representação numérica desse objeto pode considerá-lo um ponto no espaço de n dimensões.

Esse estudo é centrado no modo não-supervisionado de aprendizado, através de métodos de análise de conglomerados (*clusters*). As próximas seções detalham

aspectos referentes a essa questão, além de trazer uma breve descrição do aprendizado supervisionado.

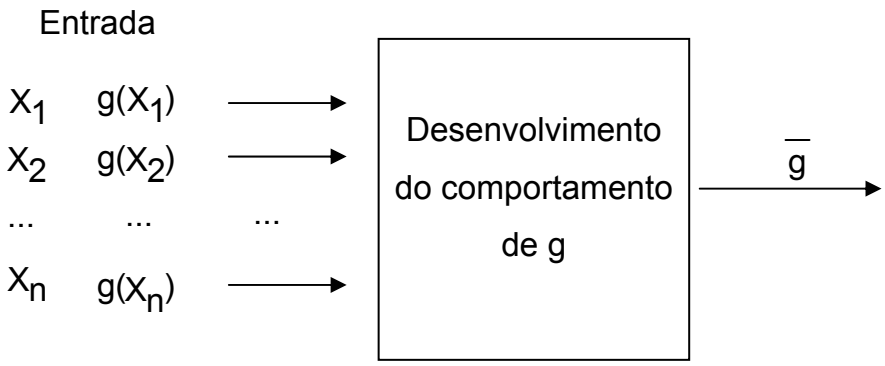
Aprendizado Supervisionado

Conforme citado anteriormente, os temas abordados por técnicas de aprendizado computacional são tipicamente de duas modalidades: classificação de objetos e resolução de problemas.

Para ambas modalidades, foi visto que o mecanismo de aprendizado pode ser caracterizado quanto ao seu modo e ainda quanto à ausência ou presença de supervisão do processo. Como o produto desse estudo será aplicado com a intenção de estabelecimento de categorias de operação dos nós da InteGrade, o foco do texto será o problema de classificação de objetos.

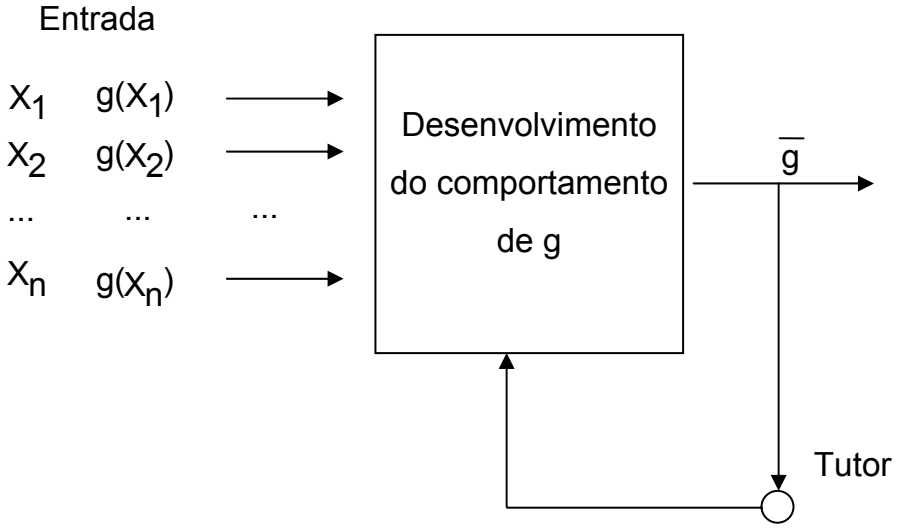
Como visto, no aprendizado supervisionado, um conjunto de objetos pré-classificados é disponibilizado. Esses objetos são descritos por um vetor de características, normalmente composto de valores numéricos. Assim, supondo a existência de *N* objetos, cada um definido por *m* características, podemos definir Xij, $1 \leq i \leq N$, $1 \leq j \leq m$, como o valor da característica *j* do objeto *i*. Como esses objetos já são pré-classificados a cada vetor $Xi = \{X_{i1}, X_{i2}, \ldots, X_{im}\}$ é associada uma categoria Wk, com $1 \leq k \leq c$, onde *c* é o total de categorias existentes.

O problema do trabalho de aprendizado passa a ser o desenvolvimento de uma função g(X) que receba como parâmetro um vetor de *m* dimensões e retorne o valor de categoria Wk, com $1 \leq k \leq c$. Na prática, é desenvolvida uma função $\overline{g}$. O objetivo é minimizar a distância entre g e $\overline{g}$, de acordo com alguma medida de distância definida. Isso pode ser mais bem descrito pelo diagrama:



Com esse procedimento foi obtido um classificador (função $\overline{g}$ que foi desenvolvida) que permite a identificação da classe à qual pertence um dado objeto m-dimensional.

A evolução desse modelo prevê o aprendizado constante, quando novos objetos classificados passam a aprimorar a medida de desempenho do classificador $\overline{g}$. Com a ajuda de um tutor ou oráculo, é feita uma retro-alimentação no diagrama anteriormente elaborado:



Cumpre-se o objetivo de otimizar a função de classificação $\overline{g}$, melhorando algum indicador de desempenho pré-estabelecido. Os mecanismos de

classificação enfocados na ausência de exemplos de treinamento e de tutores pré-definidos serão vistos a seguir.

Aprendizado Não-Supervisionado

Pode-se referir ao aprendizado não-supervisionado como aquele que se dá sem a existência de um conjunto de treinamento com exemplos pré-classificados. Sem esse corpo de treino, é necessário que inferências sejam feitas sem conhecimento *a priori* dos objetos de estudo e sem um resultado esperado já definido. Na prática, um dado de entrada deve ser avaliado e classificado sem que, para isso, existam outros objetos de referência com os quais ele possa ser comparado.

Como não há uma massa de dados de exemplo contra a qual confrontar os eventos e dados de entrada recebidos por um sistema de aprendizado não-supervisionado, a análise da redundância da informação recebida é fundamental no processo da construção do conhecimento. Essa redundância pode ser obtida na aferição de algumas medidas na massa de dados de entrada, como:

- média – alguns fenômenos podem ser analisados com a simples observação do valor médio de ocorrência de uma dada grandeza que os caracteriza, auxiliando no processo de aprendizado. Por exemplo, mesmo sem o auxílio de técnicas de meteorologia, é possível fazer uma estimativa rudimentar da temperatura climática de um lugar inexplorado observando-se a temperatura desse ambiente ao longo de um determinado tempo e daí se extraindo a média das temperaturas obtidas. Na maioria dos casos, como o próprio exemplo, o conhecimento da média não pode ser considerado determinante no processo de aprendizado computacional, mas pode ser um de seus valiosos recursos. O próprio cálculo da média está sujeito às diversas metodologias existentes tais como a média aritmética, média harmônica e a média geométrica, entre outras.
- Variância – quando a simples observação da média de uma série de dados não for suficiente para se chegar a uma conclusão definitiva

quanto à organização dos dados, pode-se fazer uso da medição de variância. A variância pode ser entendida como o grau de dispersão dos valores observados de uma dada medida. Por exemplo, a média da cotação de um papel financeiro poderia induzir a um erro de avaliação se não fosse observada a variância de cotação desse papel. Mesmo com um bom desempenho médio, uma grande variância por ser indicativa de possíveis grandes variações, negativas ou positivas, da cotação desse papel, o que o torna extremamente arriscado.

- Covariância – essa medida estabelece a correlação entre os diversos conjuntos de objetos observados. Um estudo de população de animais pode-se valer da medição de uma alta correlação entre as quantidades de animais de duas espécies distintas para inferir a população de uma das espécies baseando-se no crescimento populacional da outra espécie.
- Regras para combinação e aglomeração de objetos – outra linha de avaliação da ocorrência de redundância nos fenômenos observados é o entendimento de como os objetos se agrupam em classes de objetos com características semelhantes. Esse é o foco de uma das principais técnicas de aprendizado não-supervisionado, o agrupamento ou *clustering*, que será discutido com maior aprofundamento neste estudo.

Métodos de Agrupamento

Classificação

A classificação é um dos processos fundamentais na ciência, uma vez que os fatos e fenômenos devem ser ordenados antes de poder-se entendê-los e desenvolver-se princípios que expliquem sua ocorrência. Conceitualmente, entende-se por classificação o processo de ordenação de objetos por suas semelhanças.

Clustering

Clustering é um tipo de classificação especial, feita sem o auxílio de um corpo de treinamento. Apesar de a definição de um cluster não ser algo bem preciso,

há um entendimento genérico de que membros de um *cluster* são mais semelhantes entre si que membros externos a esse cluster. O conceito de semelhança está associado à distância entre objetos considerando-se o espaço de atributos que caracterizam um dado objeto.

Passos gerais de uma análise de conglomerados

Milligan [Mil96] propõe alguns passos básicos que devem ser seguidos numa implementação padrão de uma análise de *clustering*. Muitas vezes pode-se ter a falsa impressão de que o método escolhido representa toda a análise, mas trata-se apenas de um dos passos que devem ser seguidos. Há casos particulares em que esses passos devem ser adaptados (com a inclusão ou exclusão de algumas das etapas), mas de uma forma geral são:

1. Definição dos elementos a serem agrupados
2. Seleção das variáveis
3. Padronização e normalização das variáveis
4. Definição das medidas de similaridade ou distinção
5. Escolha do método de agrupamento
6. Número de *clusters* a serem considerados
7. Testes e interpretação dos resultados

1. Definição dos elementos a serem considerados

A seleção dos elementos que serão utilizados no processo de agrupamento é decisiva na determinação da estrutura dos *clusters* que serão criados. Afinal, toda a metodologia será aplicada sobre esses exemplos selecionados. Apesar disso, a pesquisa a respeito desse tópico de discussão ainda é bem limitada.

Uma técnica que pode ser empregada é a seleção aleatória de exemplos. Quando se pretende generalizar o resultado obtido para um grande número de elementos, pode-se tomar uma amostragem selecionada aleatoriamente da população de objetos disponíveis. Entretanto, esse método de seleção deve ser usado com bastante critério, pois traz o risco de a amostragem selecionada não ser representativa do espaço amostral pesquisado. A regra básica na escolha

dos elementos é que eles devem representar de forma consistente o conjunto de objetos existentes.

Outra possibilidade de recurso a ser utilizado é definição de tipos ideais. Durante a pesquisa e modelagem do processo de agrupamento, podem surgir alguns tipos prováveis de serem encontrados nos dados disponíveis. No caso em estudo, InteGrade, há expectativa de serem encontrados alguns comportamentos típicos entre os nós pertencentes à grade: dia normal de trabalho, dia atarefado e feriado. Cada um desses comportamentos deve pertencer a um provável *cluster* a ser encontrado, que poderá ser tipificado num elemento ideal. Esse elemento ideal é incluído na massa de dados e submetido ao processamento com os demais objetos.

Após a conclusão do processo de agrupamento, pode-se fazer verificações tais como: verificação se todos os tipos ideais ficaram em *clusters* distintos ou se surgiu algum *cluster* que não foi previsto durante a elaboração dos tipos ideais. Com os resultados obtidos, é possível fazer a remodelagem do processo.
Finalmente, pode-se também levar em consideração a presença de elementos atípicos nos dados de entrada. Isso se dá porque alguns objetos disponíveis para análise podem não pertencer a nenhum *cluster* apropriadamente, representando uma espécie de ruído de dados. Alguns métodos disponíveis na literatura, como o método de Ward [War63], são resistentes a ruídos, mas uma forma alternativa é uma intervenção manual de retirada de eventuais ruídos e uma posterior análise comparativa de validação com os *clusters* obtidos.

2. Seleção das variáveis

As variáveis escolhidas para caracterizar os objetos têm extrema importância na análise clustering. Elas que irão determinar como os objetos serão agrupados, já que a comparação entre os diversos objetos presentes na massa de dados é feita com base nas medidas dessas variáveis. Apenas as variáveis que realmente têm importância para o agrupamento a ser realizado devem ser consideradas.

Um especialista no domínio da informação sobre o qual está sendo desenvolvido o trabalho deve ser envolvido, pois a caracterização de um objeto muitas vezes depende de uma análise subjetiva. Ainda deve ser definido, também com a ajuda desse especialista, como se dará a medição da característica. As grandezas devem ser quantificadas, pois muitas vezes são grandezas qualitativas, para ficarem sujeitas aos métodos de agrupamento.

O grau de correlação entre as variáveis escolhidas deve ser levado em consideração, uma vez que variáveis bastante correlatas podem acabar dominando a caracterização de um objeto. Isso não necessariamente é ruim, mas o analista deve ter em mente que o papel que cada característica deve ter na definição de um objeto.

Outro fenômeno que merece atenção nesse processo é a ocorrência de variáveis que distorcem a caracterização dos objetos (masking variables). O uso inadequado de variáveis que não possuam uma forte justificativa para serem consideradas no processo pode levar a uma formação de grupos completamente distinta da realidade. Muitos estudos práticos já concluíram que a simples adição de uma ou duas variáveis como ruído leva à formação de clusters completamente diferentes. O uso da distância euclidiana ponderada como medição de proximidade entre os objetos pode minimizar o efeito da eventual existência desse tipo de variáveis.

3. Padronização e normalização das variáveis

Em casos onde há grande diferença entre as magnitudes médias ou entre as variâncias das medidas das variáveis, muitos analistas são levados a crer que a padronização dos dados é mandatória. A maioria dos métodos de análise de conglomerados não assume, entretanto, a existência de dados padronizados. A normalização ou a padronização, apesar de poderem ser aplicadas, nem sempre são necessárias. A existência de um cluster pode ter relação com o espaço de dados original e a padronização dos dados pode prejudicar o agrupamento dos dados.

Se a existência de clusters é assumida em um espaço de dados transformados, então a padronização poderá ser aplicada. O novo questionamento diz respeito a como proceder à padronização. Milligan e Cooper (Mil88) enumeram oito formas de fazer a medição das variáveis.

Seja x o valor medido de uma característica, $\bar{x}$ o valor médio da característica entre todas os objetos, s o desvio padrão, e Max, Min e Rank funções que retornam, respectivamente, o maior valor medido para uma característica, o menor valor medido e a posição ordinal da medida x dentro da distribuição de medidas. As medidas propostas são:

z0 = x (sem padronização)

z1 = $\frac{x - \bar{x}}{s}$

z2 = $\frac{x}{s}$

z3 = $\frac{x}{Max(x)}$

z4 = $\frac{x}{Max(x) - Min(x)}$

z5 = $\frac{x - Min(x)}{Max(x) - Min(x)}$

z6 = $\frac{x}{\sum x}$

z7 = Rank(x)

Cada uma dessas medidas pode ser aplicada em condições distintas, o que leva à necessidade de análise prévia e de realização de alguns testes para a determinação de qual método deverá ser utilizado.

4. Definição das medidas de similaridade ou distinção

A medida de similaridade entre objetos corresponde à métrica na qual se acredita que os clusters existam. O próprio entendimento de cluster envolve

essa medida, já que objetos dentro de um cluster são mais próximos (similares) entre si que os demais objetos.

Muitas vezes é mais simples medir a distinção de objetos, em vez da similaridade. Mas é imediato que os conceitos são complementares e a escolha de qual abordagem usar dependerá de como os objetos são caracterizados.

Como na maioria dos casos um objeto é caracterizado em um espaço n-dimensional, a distância euclidiana (uma medida de distinção) nesse espaço se tornou uma das medidas mais amplamente utilizadas para a proximidade entre os objetos. Essa e outras medidas são consideradas mais apropriadamente em seção específica nesse estudo.

5. Escolha do método de agrupamento

O método de agrupamento é a parte central da análise *clustering*. Há quatro aspectos [Mil96] que devem ser observados na escolha de um método:

- o método deve ser projetado de forma a recuperar os grupos que se espera estarem presentes na massa de dados;
- o método deve ser efetivo ao recuperar as estruturas para as quais foi projetado;
- o método deve ser capaz de lidar com erros presentes na massa de dados;
- deve ser possível o projeto de software que se adeque ao método.

Neste estudo, serão considerados os métodos clássicos de *clustering*: algoritmo hierárquico e algoritmo de *k-centros*. Esses métodos serão discutidos adiante.

6. Número de clusters a serem considerados

Muitos métodos de *clustering* não levam em consideração o problema de determinação do número de grupos existentes na massa de dados. Ao contrário, nesses métodos um especialista deverá indicar o número de *clusters* desejados ao final do processamento dos dados. Os métodos hierárquicos e os algoritmos

baseados no método *k-centros* assumem que os dados devem ser agrupados em um número pré-definido de classes.

A maioria das tentativas de pesquisadores na direção de determinar esse número ideal se deparou com um problema de otimização. Dada uma massa de dados, diversas iterações do algoritmo de *clustering* são feitas, na busca de um número de *clusters* que minimize alguma medida de erro determinada. Além de essas tentativas terem se mostrado extremamente caras em tempo computacional, há um problema intrínseco a essa otimização: muitas vezes o menor erro será para o caso de considerar tantos *clusters* quantos forem os objetos a classificar. Cada *cluster* conterá um único objeto, o que não é a solução desejada.

Na aplicação para a InteGrade, será adotado um número inicial de cinco grupos. Esse um número pode ser considerado razoável, uma vez que cada grupo representa um modo de operação de um determinado nó da grade. Experimentalmente, deverão ser consideradas outras alternativas, possibilitando uma análise comparativa entre as proposições.

7. Testes e interpretação dos resultados

A interpretação dos resultados e o estabelecimento de medidas de avaliação de desempenho são as últimas, mas não menos importantes, etapas da análise. A participação de um analista da área de pesquisa é mais uma vez fundamental, já que parte do esforço nesta etapa do trabalho depende também de uma análise subjetiva. A coleta de dados estatísticos sobre cada *cluster* deve auxiliar a interpretação.

Os testes dos resultados devem ser baseados em algum objetivo mensurável previamente estabelecido. Na InteGrade, como a sugestão do uso de *clustering* visa a inferência de comportamento de uso dos recursos, uma medida possível de ser realizada na avaliação e testes do método utilizado é o índice de acertos nas inferências realizadas.

Distâncias

A distância entre dois objetos é a medida de comparação de semelhança (ou diferença) entre eles. Seja Xij a valor medido para a característica *j* do objeto *i* e *n* o número de características que descrevem um objeto. A forma mais difundida de comparar dois objetos é através do cômputo da distância euclidiana:

$$d(i,k) = \sqrt[2]{\sum_{j=1}^{n} (Xij - Xkj)^2}$$

Outra abordagem aplicável é o uso de pesos para as variáveis, resultando na distância euclidiana ponderada:

$d(i,k) = \sqrt[2]{\sum_{j=1}^{n} w_j (Xij - Xkj)^2}$, onde w_j é o peso para a característica *j*. Esse peso pode ser atribuído de formas distintas:

- subjetivamente – a escolha dos pesos é feita por um especialista no momento de escolha das variáveis que farão parte da descrição de um objeto;
- escala de variância – o peso é inversamente proporcional à variância das medidas da característica.

Quando há presença de variáveis com alta correlação, pode-se minimizar o efeito dessas variáveis com o uso de uma matriz $\overline{w}_{jl}$ de relação entre as variáveis. Chega-se à definição da distância euclidiana generalizada:

$$d(i,k) = \sqrt[2]{\sum_{1 \le j,l \le n} w_{jl} (Xij - Xkj)(Xil - Xkl)}$$

Há na literatura uma outra forma de generalização da distância euclidiana, conhecida com família de Minkowski:

$d(i,k) = \sqrt[p]{\sum_{j=1}^{n} d(i,k \mid j)^p w_j}$, onde p é um parâmetro que define o elemento da família e $d(i,k|j)$ é distância entre os objetos *i* e *k*, com respeito à variável *j*.

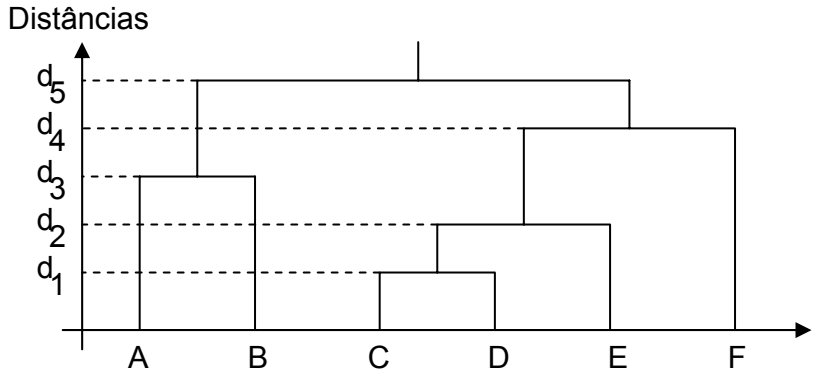
Clustering Hierárquico Aglomerativo

Esse método de análise *clustering* é bem simples:

Supondo que se deseja fazer a análise de *n* itens (objetos), é montada uma matriz de distâncias, de dimensão *n*x*n*, contendo as distâncias entre os elementos. Os dois elementos mais próximos, de menor distância, são unificados gerando uma nova matriz, de dimensões (*n*-1)x(*n*-1). A unificação dos elementos resulta na criação de um novo objeto, um grupo, contendo os elementos unificados.

Esse procedimento é repetido até que todos os dados sejam condensados num único grupo. Esses passos são guardados de forma a construir um dendograma, definido a seguir.

O dendograma é um diagrama em forma de árvore, bastante familiar aos taxonomistas. A forma geral de um dendograma é apresentada com uma raiz ao topo, com uma escala de distâncias indicando em que nível dois grupos tornam-se um só. A figura abaixo exemplifica o agrupamento sucessivo dos objetos A, B, C, D, E e F.



As distâncias d_1, d_2, d_3, d_4 e d_5 representam os pontos em que os agrupamentos foram formados. Dependendo do nível de agrupamento e quantidade de *clusters* desejada, pode-se tomar um determinado valor para se estabelecer a resposta de quais objetos estão contidos em quais clusters. Por exemplo, se tomarmos

como base um valor entre d_3 e d_4 na figura acima, haverá a presença de três clusters: {A, B}, {C, D, E} e {F}.

Duas importantes medidas tornam-se necessárias:

- a distância entre dois objetos individuais, já abordada anteriormente.
- a distância entre dois grupos de objetos

A distância entre dois grupos poderia ser pensada, a princípio, em termos das coordenadas dos objetos que compõem o grupo. Mas, a menos que a distância utilizada tenha sido a euclidiana, esse tipo de abordagem pode ser contraditório, conforme pesquisas de métricas de distâncias já realizadas. O uso das coordenadas iniciais só terá sentido no contexto dos grupos para o caso euclidiano.

O grau de diferença entre dois grupos, medido através de sua distância, sugere algum tipo de média das distâncias individuais ou cômputo das distâncias entre os centros dos conglomerados. Mas, na prática, utiliza-se uma das duas formas a seguir:

- ligação simples (single linkage) – a distância entre dois grupos é definida como a menor distância entre o membro de um grupo e um membro do outro;
- ligação completa (complete linkage) – a distância entre dois grupos é definida como a maior distância entre um membro de um grupo e um membro de outro grupo.

Esses métodos podem levar a resultados bem diferentes, mas quando há forte aglomeração de objetos, resultando na existência de *clusters* naturais, os diversos métodos devem levar a resultados similares.

Uma conseqüência imediata da ligação simples é que, quando um objeto é adicionado a um grupo, a distância entre esse grupo e os demais elementos ou diminuirá ou permanecerá a mesma. Isso implica que grandes grupos tendam a

crescer e serem unificados, enquanto pontos isolados assim permanecerão até as últimas iterações do procedimento.

Um fenômeno que contribui bastante para a popularização desse método de ligação simples é o comportamento do algoritmo para casos em que haja distâncias muito similares na massa de dados. Se, num dado momento, ocorrem distâncias muito próximas umas das outras, será natural que os objetos sejam unificados em seqüência, o que realmente é esperado num cluster. Mas, se qualquer outro método for usado, isso não necessariamente irá ocorrer.

No caso da ligação completa, ocorre o inverso: sempre que um objeto é adicionado a um grupo, sua distância aos demais objetos ou será maior ou será igual à original. Isso quer dizer que quanto mais um grupo crescer, menor a tendência dele se unir aos demais grupos. O comportamento observado é que os objetos se juntem primeiro em pequenos grupos e só então os grupos são concatenados. Isso é particularmente útil quando se deseja a simples divisão de dados em grupos, em vez da busca por *clusters* naturais.

Os efeitos colaterais dessa abordagem são:

- pontos intermediários entre *clusters* naturais podem fazê-los se unir em grupos antes do momento devido, principalmente no caso da ligação simples;
- se os dados constituírem grupos de tamanhos muito distintos, a ligação composta pode unir os grupos pequenos antes de os grandes estarem completos.

Algoritmo K-Centros

Há um grande número de combinações para a realização da tarefa de dividir um conjunto de n objetos em c grupos. As diversas técnicas de clustering objetivam fazer esse agrupamento de forma que os objetos de um mesmo grupo sejam próximos (uma vez definido esse conceito de proximidade) e que objetos de grupos distintos sejam distantes.

Tomando-se uma dada partição de dados qualquer, pode-se definir uma medida de erro entre a distribuição dos n objetos e a forma como eles foram particionados. Seja a partição Pi(n,c) de n objetos em c grupos. O erro associado a essa partição será expresso por e[Pi(n,c)]. A idéia intuitiva que guia os métodos de clustering é fazer com que se tome a partição que retorne o menor valor de erro. Essa partição será considera ótima e, portanto, representa a agrupamento de objetos desejado.

Comparar o valor da medida de erro para cada uma das partições possíveis é impraticável, mesmo com alto poder computacional. O algoritmo de K-centros apresenta uma técnica de otimização desse procedimento, através da definição de uma vizinhança de partições para cada partição. Partindo-se de uma partição inicial, é feita uma busca na vizinha de partições por aquela que a medida de erro e seja mínima. Toma-se então a partição encontrada e inicia-se o processo de iteração desses passos. Com a definição de uma regra de parada, que pode ser a situação em que o erro de toda a vizinhança é maior que o da partição atual, chega-se a uma situação ótima, mesmo que se esteja falando em otimização local.

Passos do algoritmo k-centro

Sejam n objetos, definidos por um vetor m-dimensional de atributos. Assim, o valor Xij é a medida da característica j para o objeto i, com $1 \leq i \leq n$ e $1 \leq j \leq m$. Uma partição P(n,c) é tal que cada um de seus n objetos seja alocado em um único grupo entre os possíveis 1, 2, ... ou c. A média da variável j ($1 \leq j \leq m$) para o cluster l ($1 \leq l \leq c$) é expressa por b(l,j). O número de objetos pertencentes ao cluster l ($1 \leq l \leq c$) é expresso por t(l).Determina-se então a distância d(i,l) entre o objeto i ($1 \leq i \leq n$) e o cluster l. Tomando-se por base a distância euclidiana, pode-se definir:

$$d(i,l) = \sqrt[2]{\sum_{j=1}^{m} (X_{ij} - b(l,j))^2}$$

A medida de erro para essa partição pode ser definida como a somatória dos quadrados das distâncias entre os objetos e o cluster a que ele pertence. Assim: e[P(n,c)] = $\sum_{i=1}^{n} d(i,l(i))^2$, onde l(i) é o cluster, $1 \leq l(i) \leq c$, a qual pertence o objeto *i*.

Como explicado, os passos do algoritmo consistem em procurar a partição com o menor erro e, através do remanejamento dos objetos de um cluster para outro.

Passo 1.

É assumida uma configuração inicial P(n,c) de clusters 1, 2, ..., c. São computadas as médias b(l,j), $1 \leq l \leq c$, $1 \leq j \leq m$, das diversas características para cada cluster. Daí, é obtido o erro da partição, e[P(n,c)].

Passo 2.

Para o primeiro objeto, calcule a diferença no erro causada pelo remanejo do primeiro objeto i=1 do grupo l(i) para cada um dos outros c-1 grupos restantes.

A diferença no erro é expressa por:

$\frac{n(l)d(i,l)^2}{n(l)+1} - \frac{n(l(i))d(i,l(i))^2}{n(l(i))-1}$, para a transferência do referido objeto para o grupo l, $1 \leq l \leq c$.

Se o valor mínimo dessa diferença, considerando todos os possíveis valores de $l \neq l(i)$, for negativo, mova o objeto i=1 de l(i) para l, recalculando o centro dos 2 grupos e ajustando o erro e[P(n,c)] com o valor da diferença encontrada.

Passo 3.

Repita o passo 2 para todos os demais objetos ($2 \leq i \leq n$).

Passo 4.

Se não houve movimento algum de qualquer dos objetos, pare. Se houve, repita o passo 2.

Outras distâncias aplicáveis

A característica essencial do algoritmo K-centros é a busca pela otimização local, através do remanejo de objetos entre os clusters, e a medida de distância. Como se fala em centro de cluster, é natural que o conceito utilizado seja o da distância euclidiana.

Entretanto, pode-se utilizar outros conceitos de distâncias, com a devida extensão das definições envolvidas no método. Seja o objeto i denotado por Xi, $1 \leq i \leq n$, e b(l)={b(l,j), j=1,...,m} o conjunto de valores correspondentes ao cluster l. A distância F[Xi,Xk] diz respeito a distância entre dois objetos i e k. Dessa forma, o objeto central do cluster l é b(l) que minimize a expressão $\sum_{i \in l} F(X_i, b(l))$.

Conclusão e considerações finais

Com esse estudo, foi possível desenvolver o embasamento teórico necessário para a aplicação ao estudo de caso de *clustering* de dados no âmbito do projeto da InteGrade. Particularmente, são sugeridos dois métodos que poderão ser utilizados: algoritmo *k-centros* e algoritmo hierárquico.

A evolução desse estudo prevê a aplicação desses métodos à massa de dados já coletada para agrupamento e posterior análise dos resultados do uso desse agrupamento na inferência do comportamento de uso dos recursos. Variações dos métodos propostos ou aproximações existentes na literatura deverão ser pesquisadas e utilizadas para efeitos de comparação com o que aqui foi proposto.

Bibliografia

(Fuk90] Fukunaga, K. Introduction to Statistical Pattern Recognition. Academic Press Inc., 2. ed., 1990.

[Har95] Hartigan, J.A. Clustering Algorithms. John Wiley & Sons, 1975.

[Krm95] Krzanowski, W.J. & Marriot, F.H.C. Multivariate Analysis Part 2: Classification, covariance structures and repeated measurements. Arnold, 1995.

[Lan96] Langley, P. Elements of Machine Learning. Morgan Kaufmann Publishers, Inc, 1996.

[Mil88] Milligan, G.W. & Cooper, M.C. A study of variable standardization. Journal of Classification, 5, 181-204. 1988.

[Mil96] Milligan, G.W. Clustering validation: Results and implications for applied analyses. Clustering and Classification. P. Arabie, L.J.Hubert & G. De Soete (Eds.), World Scientific Publishing, 1996.

[Sok76] Sokal, R.R. Clustering and Classification: Background and Current Directions. In Proceedings of the Advanced Seminar on Classification and Clustering, 1976.