



**COPPE/UFRJ**

UTILIZAÇÃO DE TÉCNICAS DE DATA MINING NA DETECÇÃO DE OUTLIERS EM  
AUXÍLIO À AUDITORIA OPERACIONAL COM UM ESTUDO DE CASO COM DADOS  
DO SISTEMA DE INFORMAÇÕES HOSPITALARES

Antonio Carlos Bodini Junior

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Civil.

Orientador: Alexandre Gonçalves Evsukoff

Rio de Janeiro  
Setembro de 2009

UTILIZAÇÃO DE TÉCNICAS DE DATA MINING NA DETECÇÃO DE OUTLIERS EM  
AUXÍLIO À AUDITORIA OPERACIONAL COM UM ESTUDO DE CASO COM DADOS  
DO SISTEMA DE INFORMAÇÕES HOSPITALARES

Antonio Carlos Bodini Junior

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ  
COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA  
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS  
NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM  
ENGENHARIA CIVIL.

Aprovada por:

---

Prof. Alexandre Gonçalves Evsukoff, Dr.

---

Prof. Nelson Francisco Favilla Ebecken, D.SC.

---

Prof. Beatriz de Souza Leite Pires Lima, D.Sc.

---

Prof. Elton Fernandes, D. Sc.

---

Prof. Marley Maria Bernardes Rebuszi Vellasco, Ph. D.

RIO DE JANEIRO, RJ - BRASIL  
SETEMBRO DE 2009

Bodini Junior, Antonio Carlos

Utilização de técnicas de data mining na detecção de *outliers* em auxílio à auditoria operacional com um estudo de caso com dados do Sistema de Informações Hospitalares/ Antonio Carlos Bodini Junior. – Rio de Janeiro: UFRJ/COPPE, 2009.

VIII, 122 p.: il.; 29,7 cm.

Orientador: Alexandre Gonçalves Evsukoff

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia Civil, 2007.

Referencias Bibliográficas: p. 110-119.

1. Mineração de Dados. 2. Agrupamento Nebuloso. 3. Dados anômalos. I. Evsukoff, Alexandre Gonçalves. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

## **Agradecimentos**

A pergunta mais comum que se ouve durante todo o período de estudo é: qual a sua contribuição com esse trabalho? Sobre a contribuição, acredito que o próprio trabalho responde a esta questão. Entretanto, sobre o pronome possessivo, julgo estar errado, devendo-se trocar o “minha” por “nossa”, pois isto é fruto de várias mãos, que direta e indiretamente auxiliaram-me, compartilhando surpresas e sobressaltos e provendo idéias. Assim, cabe agradecer.

Primeiramente ao Pai eterno, pela graça da vida e pela dádiva dessa aventura na Ilha do Fundão.

À Professora Ilara Hammerli Sozzi de Moraes, agradeço o auxílio desinteressado, que, mobilizando meios na Secretaria de Estado de Saúde para prover-me de dados e informações, possibilitou a concretização de uma idéia.

Ao Professor Alexandre Gonçalves Evsukoff, meus agradecimentos pela forma segura e sempre bem humorada como conduziu a orientação. Sem sua ajuda, o objetivo não seria atingido. Agradeço o envio constante de artigos técnicos, disponibilização de livros e, sobretudo, as idéias sobre o ferramental adequado ao estudo. Largo em afazeres, sempre teve tempo para pensar, não apenas em como orientar, mas como resolver o problema. Mais que um orientador, um amigo!

Ao Professor Nelson Francisco Favilla Ebecken, agradeço por ter me acolhido neste programa. Participar de suas aulas foi mais que aprender, foi viajar pelo universo da mineração de dados, um estímulo a buscar novos horizontes. Sua participação na banca é a certeza do aprimoramento do trabalho, face às suas abordagens práticas amparadas na experiência e conhecimento.

À Professora Beatriz de Souza Leite Pires Lima, tendo a sorte de participar de suas aulas de métodos bio-inspirados, agradeço por ter apresentado um mundo novo no universo dos algoritmos, onde sua forma de apresentação faz com que os alunos se apaixonem pelo assunto. Obrigado, também, pelas observações acerca do trabalho, que o lapidam, tirando-o do estado bruto.

Agradeço ao Professor Elton Fernandes pela participação na banca, cuja ótica da Engenharia de Produção muito contribui para melhoria do trabalho.

À Professora Marley Maria Bernardes Rebutzi Vellasco, meus agradecimentos pela aceitação em participar da banca mesmo com uma agenda

apertada e pelas observações acerca do trabalho, que são determinantes para o melhoramento do mesmo.

Agradeço aos componentes do programa de Engenharia Civil, que suportam e solucionam os problemas gerados pelos alunos. Agradeço à Egna, ao Jairo, à Ana, Beth, Célio e Orlando por estarem sempre presentes.

Um agradecimento especial à Solange Coelho de Oliveira, cuja palavra amiga foi essencial para restituir-me a calma de espírito e permitir o prosseguimento nas últimas etapas desse projeto.

Agradeço os amigos de trabalho e de paróquia, que torceram para que tudo desse certo.

Finalmente, à minha retaguarda, guarnecida por minha esposa, cúmplice de todos os momentos e em todos os assuntos, filhos, torcedores fanáticos de meu sucesso, e mãe e tia, companheiras de ansiedade, agradeço as orações e pensamentos para que tudo viesse a frutificar.

A todos, nas palavras de Francisco, Paz e Bem!

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

UTILIZAÇÃO DE TÉCNICAS DE DATA MINING NA DETECÇÃO DE OUTLIERS EM AUXÍLIO À AUDITORIA OPERACIONAL COM UM ESTUDO DE CASO COM DADOS DO SISTEMA DE INFORMAÇÕES HOSPITALARES

Antonio Carlos Bodini Junior

Setembro/2009

Orientador: Alexandre Gonçalves Evsukoff

Programa: Engenharia Civil

Uma auditoria busca verificar ocorrência de um desvio nas lides administrativas e, se possível, as razões de sua ocorrência. Os desvios são caracterizados pelo registro de dados anômalos e sua evidenciação em uma base de informações não é fácil, mormente em sistemas de grande volume, e fica subordinada à experiência do auditor. Buscando facilitar a descoberta destes, propõe-se o uso de técnicas de mineração de dados no auxílio de auditagens por intermédio dos algoritmos de Agrupamento Nebuloso (*Kernel Possiblistic C-Means*), Máquina de Vetor Suporte (*SVM – One Class*), que permitem a evidenciação dos *outliers*. Desse estudo, verificou-se o uso viável de um novo algoritmo com o mesmo fim, a Função de Similaridade Média, mais simples e igualmente eficaz nesse objetivo. Submeteram-se os dados do Sistema de Informações Hospitalares aos algoritmos, o que demonstrou a efetividade desses na descoberta de *outliers* e comprovou a utilidade do novo algoritmo.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

USE OF DATA MINING METHODS TO DETECT OUTLIERS TO AID OPERATIONAL  
AUDIT WITH CASE STUDY IN HEALTH INFORMATION SYSTEM'S DATA

Antonio Carlos Bodini Junior

September /2009

Advisor: Alexandre Gonçalves Evsukoff

Department: Civil Engineering

An audit search in administrative works and, if possible, the reasons for it. It is very hard to discover such errors in large data bases and the discovery is normally based in the analyst's experience. To make it easier, we propose the use of data mining techniques in order to help auditings: a fuzzy clustering algorithm, Kernel Possibilistic C-Means, and a support vector machine method, SVM One Class. These algorithms are able to discover outliers in a dataset. From this research, it was verified a viability of the use of a new algorithm with the same finality, Mean Similarity Function. The effectiveness of the use of these algorithms was tested in Health Information System's data in which were discovered outliers.

## Conteúdo

1	Introdução .....	1
2.	Metodologia Aplicada .....	6
2.1	Trabalhos relacionados .....	6
2.2	Detecção de dados anômalos .....	9
2.2.1	Métodos estatísticos .....	10
2.2.2	Métodos baseados em distâncias .....	15
2.2.3	Métodos baseados em similaridade .....	26
3.	Aplicativos Desenvolvidos .....	37
3.1	Aplicativo de Agrupamento Nebuloso ( <i>KPCM</i> ) .....	37
3.2	Máquina de Vetor Suporte – Classe Única ( <i>SVM – One Class</i> ) .....	38
3.3	Aplicativo Função de Similaridade Média ( <i>FSM</i> ) .....	41
4.	Avaliação dos algoritmos .....	43
4.1	Avaliação do aplicativo de Agrupamento Nebuloso .....	43
4.2	Avaliação do aplicativo <i>SVM-One Class</i> .....	48
4.3	Avaliação do aplicativo de Função de Distância Média .....	51
4.4	Precisão dos Algoritmos .....	53
4.4.1	Precisão do aplicativo de Agrupamento Nebuloso ( <i>KPCM</i> ) .....	55
4.4.2	Precisão do aplicativo <i>SVM-One Class</i> do aplicativo <i>LIBSVM</i> .....	60
4.4.3	Precisão do aplicativo Função de Similaridade Média .....	66
4.5	Comparação com outros resultados .....	70
4.6	Considerações .....	72
5.	Descrição da Base de Dados .....	74
5.1	Atributos da Base de Dados. ....	74
5.2	Análise Exploratória dos Dados .....	75
6.	Resultados .....	83
6.1	Agrupamento Nebuloso ( <i>KPCM</i> ) .....	83
6.2	Aplicação de Máquina de Vetor Suporte com classe singular ( <i>SVM – One Class</i> ). ....	87
6.3	Aplicativo de Função de Similaridade Média .....	90
6.4	DISCUSSÃO .....	96
7.	Conclusão e sugestão para futuros trabalhos .....	105
	Referências Bibliograficas .....	110
	Anexo A – Lista de Atributos do Arquivo de Autorização de Internação Hospitalar .....	120

## 1 Introdução

São múltiplos os fins que a atividade humana procura alcançar, mas os recursos para tanto são limitados e escassos, o que impõe escolhas que maximizem sua utilidade (ROSSETI, J.P., 1994).

Os serviços de saúde pública não são exceção à regra e, coerente com esta otimização, agregou-se todos os serviços estatais - das esferas federal, estadual e municipal - e os serviços privados no Sistema Único de Saúde (SUS) em atendimento à Lei Nº 8.080/90 que regulamentou o estabelecido pela Constituição Federal de 1988. No sentido de se resguardar os recursos colocados nas entidades públicas, criou-se a Lei de Responsabilidade Fiscal, que estabelece normas de finanças públicas voltadas para a responsabilidade na gestão fiscal, mediante ações em que se previnam riscos e corrijam os desvios capazes de afetar o equilíbrio das contas públicas, destacando-se o planejamento, o controle, a transparência e a responsabilização, como premissas básicas.

A própria Constituição Federal também expõe tal preocupação, quando no artigo 37 elenca o princípio da eficiência como tônica da administração gerencial do estado, que se traduz pela qualidade da prestação do serviço à universalidade de sujeitos e de interesses e contenção de gastos desnecessários (PROCHNOW, J. J., 2000).

Assim, consequência natural, é verificar-se o atingimento dos destinos escolhidos sem o desperdício de recursos por meio de auditagens das ações dos diversos gestores, onde se visa ir além da aderência dos demonstrativos contábeis aos padrões oficiais de contabilidade, buscando avaliar a eficiência e eficácia das operações das entidades públicas e privadas, verificando se a entidade adquire, protege e usa seus recursos sem desperdícios e as causas de eventuais práticas errôneas, além de observar o cumprimento das normas legais. Estas ações não constituem medidas inovadoras, visto que a Norma de Operação Básica do SUS, em 1996, determinava que as três entidades públicas gestoras seriam responsáveis por efetuar auditorias analítica e operacional, bem como, definir os recursos e a metodologia adequada de trabalho, instrumentos para a realização das atividades, consolidar as informações necessárias, analisar os resultados e propor medidas corretivas no processo de decisão da alocação dos recursos.

Ainda que exigências de *accountability* no setor público imponham às entidades governamentais encarregadas do controle a obrigação de examinar as finanças públicas sob inúmeras perspectivas, não apenas limitadas à confiabilidade

dos demonstrativos contábeis padronizados, mas estendendo-se a todos os aspectos de gestão financeira (BITTENCOURT, F. M. R., 2005), alguns analistas vêem que há um controle fraco da gestão orçamentária e financeira (GRATERON, I. R. G, 1999).

Tal situação demanda a criação de sistemas de informação em saúde, focalizando a geração de informações e o acesso aos dados estatísticos, orçamentários, gerenciais e financeiros do Sistema Único de Saúde (SUS), bem como a existência de uma política norteando pesquisa e desenvolvimento de soluções na área da informática para aumentar a capacidade metodológica, tecnológica e gerencial, evitando que esta área seja mero depósito de dados ( VASCONCELOS, M. M., 2002).

A preocupação com o uso eficiente dos recursos está presente também no Manual de Auditoria do SUS, onde se prevê o objetivo de viabilizar a racionalização de gastos, evitando e detectando fraudes e malversação de recursos públicos. Tal objetivo, entretanto, não será alcançado pelos métodos usuais de auditoria, posto a dificuldade em fornecer dados no tempo oportuno, em virtude do volume transacionado a examinar.

Estes dados a serem analisados pelo auditor constituem *outliers*, sendo definidos como aqueles registros que se destacam, em um conjunto de dados, dos outros elementos por sua dissimilitude ou inconsistência. São objetos que não obedecem ao comportamento geral dos dados, isto é, é uma observação que desvia bastante de outras observações, levantando dúvidas acerca de sua correção (PEARSON, R. K., 2002) (XI, J., 2008). São observações que têm distribuições de probabilidade diferentes da maioria dos dados do conjunto (SCHWERTMAN *et al.*, 2004). Sua existência pode ser confundida como registro colocado erradamente (ruído) e que, portanto, deve ser expurgado da base, sob pena de degradar qualquer conclusão acerca destes (ANGIULLI, F.; PIZZUTI, C.; 2005), bem como têm grande influência na maioria dos testes paramétricos. O trato desses dados revela-se difícil, pois podem ser realmente enganos ou representar um fenômeno de interesse.

A pesquisa de *outlier* tem se demonstrado como uma área importante dentro da mineração de dados com inúmeras aplicações (como detecção de fraudes em cartões de crédito, comércio eletrônico entre outros). A detecção deste constitui um problema desafiador, tendo em vista que a razão entre estes dados e os normais usualmente é pequena (BRAUSE, R. *et al.*, 1999) (LATECKI, L. J. *et al.*, 2007) .

Conforme FILZMOSER, P. *et al.* (2008), os modelos clássicos estatísticos aplicados cegamente em dados contendo dados divergentes levam, normalmente, a resultados enganosos, pois ferramentas clássicas baseadas em média e variância são raramente capazes de detectar *ouliers*.

Diversos autores (PAPADIMITRIOU, S. *et al.*, 2002; KIRKOS, E. *et al.*, 2006) têm se dedicado a verificar um caminho mais efetivo na identificação dos dados anômalos e, independente do foco, todos perpassam pelos conceitos de agrupamento e de *ouliers*, auxiliados das técnicas estatísticas e pela teoria de conjuntos nebulosos.

Constata-se que a execução de auditorias tem se tornado difícil em face da quantidade de registros armazenados em uma organização. Para sobrepujar esta dificuldade, os auditores fazem uso de técnicas de amostragem em busca de redução da quantidade de itens a serem verificados, sem prejuízo na análise final. O desafio, no entanto, é elencar a amostra que melhor se ajusta às necessidades da auditoria, que fica normalmente subordinada à experiência do auditor (CUNHA, P. R.; BEUREN, I. M., 2006).

Conforme CUPERTINO, C. M.; MARTINEZ, A. L. (2007) outliers são candidatos naturais para constituírem o conjunto de análise do auditor e, neste sentido, as técnicas de mineração de dados são valiosas no estabelecimento destes, mormente quando não existe um conhecimento pré-determinado daquilo que seja um resultado interessante, bem como na identificação de fatores que sinalizem fatos anômalos (KIRKOS, E. *et al.*, 2007) (CALDERON, T. G. *et al.*, 2002).

O Sistema de Saúde disponibiliza os registros administrativos como forma de prover transparência de suas ações, podendo-se citar:

- Sistema de Informação de Mortalidade (SIM);
- Sistema de Nascidos Vivos (SINASC);
- Sistema de Informação Hospitalar (SIH/SUS);
- Sistema de Informação Ambulatorial (SIA/SUS);
- Sistema de Movimento de Autorização de Internação Hospitalar (AIH/SUS).

Centrando-se neste último, verifica-se lançamento de aproximadamente trinta mil registros em um ano apenas para o estado do Rio de Janeiro e, em face da demanda crescente pelos serviços públicos, a previsão de crescimento desta taxa não pode ser considerada como desatino.

Assim, qualquer auditoria sobre estes registros encontrará como primeiro óbice a escolha dos dados a serem verificados. Face a impossibilidade de analisar o todo, o auditor deve pinçar uma amostra que seja representativa da população e nesta focalizar seus esforços, traduzindo, desta forma, a confiabilidade admitida ou erro permitido (KROENKE, A. *et al.*, 2008).

Ora, esta auditoria, utilizando procedimentos normais, encontra muita dificuldade para atingir seu objetivo de elencar dados anômalos face ao desconhecimento acerca da forma como estes ocorrem e em virtude da sua falta de frequência, que impede o agregar de experiência. Estas limitações sugerem a necessidade de procedimentos mais eficazes na detecção dos *outliers* como o uso de técnicas de mineração de dados. (KIRKOS, E. *et al.*, 2007) (YUE, D. *et al.*, 2007).

Assim, consciente das dificuldades na obtenção do subconjunto de dados que devem ser analisados por um auditor e da aplicabilidade das técnicas de mineração de dados, mormente o agrupamento, como ferramenta auxiliar na auditoria, propôs-se o objetivo de comprovar a efetividade das técnicas por meio da submissão de dados reais aos algoritmos de agrupamento nebuloso e classificação de classe única. Para tanto, optou-se, respectivamente, pelas técnicas de *Kernel Possibilistic Fuzzy C-Means* (KPCM) e *Support Vector Machine - One Class* (SVM - One Class) em virtude de:

- terem a capacidade de estabelecer fronteiras não lineares entre os grupos de dados;
- o KPCM, por admitir pertinências parciais a todos os grupos, estabelece um indicativo robusto na evidenciação dos dados anômalos;
- o SVM - One Class, por ter uma alta precisão e constituir um método largamente usado em tarefas de detecção de *outliers*.

Normalmente, o estabelecimento da abrangência de uma auditoria é determinado por regras internas, que procuram determinar as ações frente a uma ocorrência, como, por exemplo, em sistema de pagamento de pessoal a quantidade de remunerações a serem verificadas em uma unidade organizacional é determinada pela quantidade de pessoas existentes naquela entidade. Pode-se também observar o uso da regra de que tudo que superar a soma da média e o triplo do desvio padrão como norma para estabelecer a fronteira dos dados normais e aqueles que devem ser auditados. Em todos estes casos não se leva a conta o fato de que, mais que

volumosos, os dados são multidimensionais, cujos atributos não devem ser vistos separadamente.

Desta forma, para não se perder a ótica multidimensional, as técnicas anteriormente mencionadas constituem um ferramental útil para um auditor decidir quais dados devem ser investigados.

Tais técnicas são eficazes, mas demandam a introdução e o entendimento de parâmetros que possibilitam o estabelecimento do subconjunto de dados a serem verificados. Assim, evidencia-se a necessidade da aplicação de uma técnica simples, porém eficaz, na elucidação do *outlier*, onde o auditor possa obter um resultado semelhante, sem ter conhecimento prévio algum acerca dos parâmetros do algoritmo. Neste rumo, depreendeu-se a viabilidade de um procedimento com esta característica, capaz de separar dados anômalos dos normais e estabelecer fronteira não linear entre esses. Esse algoritmo, decorrente do estudo aqui desenvolvido, denominou-se Função de Similaridade Média, que agrega simplicidade com capacidade de, não somente evidenciar *outliers*, como também de ordená-los segundo seu maior ou menor grau de anormalidade. Nessa função, cada registro recebe um índice indicativo de sua similitude com o conjunto de dados, que permite estabelecer quão coeso está o dado individual está em relação ao subconjunto dos dados normais.

Para tanto, este trabalho está organizado da seguinte forma:

- O capítulo 2 aborda a aplicação dessas técnicas em diversos ambientes com o objetivo de evidenciar dados anômalos e apresenta aquelas que serão utilizadas neste estudo;
- O capítulo 3 apresenta os aplicativos desenvolvidos que implementam as técnicas eleitas;
- O capítulo 4 demonstra a precisão dos algoritmos;
- O capítulo 5 descreve os dados reais a serem submetidos aos algoritmos e os caracteriza estatisticamente;
- O capítulo 6 evidencia a submissão dos dados aos algoritmos, seus resultados e apresenta uma discussão acerca da sua efetividade;
- O capítulo 7 conclui sobre a adequabilidade, exequibilidade e aceitabilidade das técnicas elencadas e sugere possibilidades para futuras pesquisas.

## **2. Metodologia Aplicada**

Mineração de dados constitui um processo válido para extração de informações úteis, previamente desconhecidas, que serão utilizadas na tomada de decisão.

Nesse processo, encontra-se observações inconsistentes em relação aos dados restantes denominadas como *outliers* e sua evidenciação demanda o estabelecimento de medidas que caracterizem explicitamente as diferenças entre estes e os normais, envolvendo um largo espectro de técnicas. Seu estabelecimento pode levar à descoberta de um novo conhecimento extremamente útil com aplicações práticas em diversas áreas como se observa nas pesquisas mencionadas nos itens seguintes. Sua determinação é imperativa em tarefas como monitoração de cartões de crédito, pagamento de seguro social, onde uma mudança de padrão pode indicar um uso fraudulento, conforme exposto por HODGE, V. e AUSTIN, J. (2004) e LU, C. *et al.*( 2003).

### **2.1 Trabalhos relacionados**

Para toda organização que transaciona milhões de registros, a análise manual torna-se inviável, posto que normalmente requer-se o resultado em prazos exíguos. Tais dados devem, portanto, ser sumarizados e, a partir desta generalização, investigadas as informações que divergem da maioria. Observa-se que a maioria das pesquisas acerca da evidenciação dos dados anômalos centra-se no mercado de crédito, seguro de saúde, de colheitas e de automóveis e telecomunicações, quase que todas restritas às áreas acadêmicas e aplicando técnicas estatísticas clássicas e métodos supervisionados e semi-supervisionados de mineração (ORTEGA, P. A., 2006)( KOU, Y., 2004).

Na busca pelos dados anômalos, podem-se agrupar dados e verificar aqueles que participam destes grupos. Estabelecer grupos de dados é uma forma de determinar diversos subconjuntos onde as distâncias entre os elementos de um subconjunto são mínimas e as distâncias entre os diversos subconjuntos são máximas. Objetiva-se agrupar dados em classes desconhecidas, utilizando medições de similaridade baseadas em distâncias entre um centro escolhido e o objeto a agrupar (YUFENG, K. *et al.*, 2004)( CHEN, Z. *et al.*, 2003), com vistas a maximizar a similaridade intra-classe (os dados de um agrupamento são semelhantes entre si) e

minimizar a similaridade inter-classes (os diversos grupos não guardam semelhanças) ( HAN, J.; KAMBER, M., 2001).

A detecção dos dados anômalos é uma preocupação clássica e uma área de pesquisa de grande interesse, como demonstram HE, Z. *et al.* (2003), quando propõem a aplicação do fator de desvio multi-granular (MDEF – *multi-granularity deviation factor*) como meio para evidenciação do *outlier*. No mesmo sentido, REN, D. *et al.* (2004) propugnam a adoção de uma medida de densidade, o fator de densidade relativa (*relative density factor* - RDF), como medida de divergência. HE, Z. *et al.* (2004) apresentam outra abordagem, onde se busca a detecção de classe de *outlier*, qual seja, um conjunto de observações agrupadas, que são anormais em relação aos outros grupos e, generalizando o conceito, passam a analisar o grau de desvio em relação à própria classe e em relação às outras classes. Para tanto, estabelecem o fator de desvio em relação à classe local (*local class outlier factor* – LCOF), que mede o grau com que um dado diverge da sua própria classe, uma medida de suspeição intra-classe, e o fator de desvio em relação a uma classe referenciada (*reference class outlier factor* – RCOF), que mede o grau de divergência de um dado em relação a uma classe em particular, uma suspeição extra-classe. A primeira medida tenta expor um dado que é divergente e a segunda, dados que guardam semelhança com os divergentes.

Ainda que tais pesquisas residam em ambientes acadêmicos, sua aplicabilidade supera estes limites, como informam CHEN, H. *et al.* (2004) ao expor que agências federais dos Estados Unidos estão envidando esforços para monitorar atividades ilegais em suas jurisdições. Envolvendo-se com a dificuldade de analisar um volume de dados crescente, concluíram que as técnicas de detecção de *outliers* poderiam ser utilizadas. Da mesma forma, LITTLE, B. *et al.* (2002) informam que ferramentas e técnicas de mineração de dados têm sido particularmente úteis na análise das faturas médicas dos seguros de saúde, onde pôde se verificar grupos que destoavam da norma, as quais foram constatadas como faturas falsas.

As ocorrências contra seguros de toda ordem são estimadas em uma perda superior a oitenta bilhões de dólares ao ano nos Estados Unidos, o que por si só determina a criação de ferramentas que as evitem e, apesar disso, observa-se que as verificações em seguros de automóveis são feitas por especialistas, onde a verificação de uma requisição indevida depende muito da experiência deste. Tal fato começa a mudar face ao estabelecimento de uma base de dados eletrônica, que permite a análise destes por meio de ferramenta semi ou totalmente automática, podendo-se

mesmo aplicar conjuntamente as técnicas de classificação como redes neurais, bayesianas e algoritmo C4.5, de forma a aproveitar as suas melhores características (PHUA, C. *et al.*, 2004) (VIAENE, S. *et al.*, 2004) (SHAO, H. *et al.*, 2002).

O sistema de saúde pública australiano, que gerencia, entre outros procedimentos, o pagamento de análises laboratoriais, tem a preocupação de evitar o pagamento de procedimentos desnecessários, excessivos ou inapropriados (ou mesmo fraudulentos), além do estabelecimento de políticas de controle. As informações registradas no sistema a uma taxa de mais de 4 milhões de registros ao ano inviabilizam qualquer análise manual. Para solucionar, buscou-se o uso de técnicas de *data mining*, tendo-se optado por agrupar as informações dos diversos laboratórios em clusters e, desta forma, verificar a existência de dados divergentes (HAWKINS, S. *et al.*, 2001).

A preocupação por assegurar a correta operação do sistema de saúde também é observada no controle de autorizações de procedimentos das entidades privadas e públicas supervisionadas pelo Fundo Nacional Chileno, que visa garantir todos os contratos, prover transparência de mercado e aumentar o conhecimento de todos os envolvidos. Em um estudo de uma empresa privada com mais de 600 mil contratantes e 18 mil médicos associados, verificou-se que a auditoria executada por dois especialistas independentes não era efetiva na prevenção da ocorrência de pagamentos indevidos, demandando, assim, outras abordagens, entre elas o uso de técnicas de mineração de dados. Tendo a oportunidade de examinar 169 requisições abusivas em uma base de 500 mil registros, pôde-se estabelecer padrões para aquelas requisições e optar-se pela construção de classificador baseado em uma rede neural com múltiplas camadas, que contribuiu com a redução de 10% nos pagamentos efetuados e uma redução de seis meses no esforço de auditoria das requisições (ORTEGA, P. A. *et al.*, 2006).

Segundo PENG, Y. *et al.* (2006), o aumento do volume de dados impedem que técnicas tradicionais (estabelecimento de regras e análise manual) de detecção de *outliers* percebam uma grande parte dessas ocorrências. Verifica-se, também, que a utilização de linguagens nativas de banco de dados do padrão *Structured Query Language* (SQL) são incapazes de elencar os dados anômalos. Assim, buscando descobrir uma ferramenta capaz de sobrepor a este óbice, analistas de uma empresa norte-americana aplicaram técnicas de *clusterização* em uma base de dados com cerca de dois milhões de registros utilizando dois aplicativos (CLUTO e SAS Enterprise Miner), com o objetivo de exibir grupos onde haveria dados suspeitos e,

dessa forma, diminuir a abrangência da análise do especialista. A premissa do método, a necessidade de se informar o número de clusters desejado, foi ultrapassada por meio de experiências e discussão com especialistas.

A preocupação com a correção da aplicação dos recursos aparece também em pesquisa no sistema de seguro de saúde de Taiwan, onde se verificava uma expansão exagerada de pacientes com doenças crônicas. Demonstrou-se diversas formas para exporem estes casos, entre eles técnicas de mineração como regressão logística, redes neurais e classificadores em árvores. Estas, entretanto, têm a limitação da falta de dados reconhecidamente como divergentes, o que dificulta a tarefa de prever a classe de um dado novo (PENG, Y. *et al.*, 2006) (LIOU, F. *et al.*, 2008).

O uso de ferramentas e técnicas de mineração de dados na detecção de registros anômalos também é amparado por resultados experimentais obtidos em grandes bases de dados de companhias de seguros, mormente aquelas nos EUA, onde pôde-se elencar diversas solicitações de pagamentos que excediam os dados normais. Neste mister, a maioria dos artigos informam que estas ferramentas, ao indicarem os dados estranhos, economizam o esforço investigativo dos analistas envolvidos (LITTLE, B. *et al.*, 2002).

Pesquisas evidenciam que a busca por estes dados divergentes não é trivial e é uma tarefa que mesmo Sherlock Holmes não seria capaz de resolver, apesar de este considerar que “há uma semelhança familiar entre todos os delitos e, tendo-se os detalhes de mil, pode-se desvendar o milésimo primeiro”. Tal assertiva não pode ser utilizada na pesquisa em grandes bases de dados, onde milhares de registros são acrescentados diariamente. A solução deste problema perpassa a automação via computador e adquire uma nuance: a separação de dados suspeitos e não suspeitos envolve um critério incerto não alcançado pelas normas usuais, necessitando-se do uso da lógica nebulosa (*fuzzy*) (BENTLEY, P., 2000a) (BENTLEY, P. *et al.*, 2000b).

## **2.2 Detecção de dados anômalos**

Autores diversos (HE, Z. *et al.*, 2004; JIN, W. *et al.*, 2004; PHUA, C. *et al.*, 2004) buscam estabelecer categorias de abordagens de pesquisa dos dados anômalos, como aquelas baseadas em distribuições de probabilidade, em profundidade, desvio entre outras. Aqui, entretanto, resumir-se-á a comparar aquelas colocadas como baseadas em modelos estatísticos clássicos (uma vez que fornecem

uma primeira descrição dos dados), e em métricas (distância e similaridade, que abrigam os algoritmos selecionados para o estudo).

Apesar de transparente, é importante observar a influência da escala, quando da análise de valores, independente da abordagem a ser seguida. Os valores das variáveis podem diferir em magnitude e/ou unidade. Algumas podem ter pesos maiores que outras na aferição do resultado e, para modificar estas influências, um pré-tratamento nos dados é mandatório (YANG, J. *et al.*, 2007). Toda análise pode levar a conclusões enganosas se não se levar em conta a escala de medida das variáveis, mormente quando se utiliza como razão de semelhança a distância entre os pontos.

Para evitar tal ocorrência, vários autores (HAN, J. e KAMBER, M., 2001; BERRUETA, L. A. *et al.*, 2007; MINGOTI, S. A, 2005, SHALABI, L. A.; SHAABAM, Z., 2006), recomendam que as variáveis sejam padronizadas por algum procedimento que diminua esta discrepância, como:

- padronização, onde o valor de cada variável é subtraído da média e dividida pelo desvio padrão:

$$\text{Valor normalizado} = \frac{\text{Valor original} - \text{Média dos valores originais}}{\text{Desvio padrão dos valores originais}}$$

- auto-escalamento, onde se opera uma transformação linear sobre os dados originais. Sabendo-se que  $Min_o$  e  $Max_o$  são, respectivamente, os valores mínimo e máximo de um atributo do conjunto de dados, e considerando que se deseja levá-los a uma escala  $[NovoMin, NovoMax]$ , mapeia-se os valores segundo:

$$\text{Valor padronizado} = \left( \frac{\text{Valor original} - Min_o}{Max_o - Min_o} \right) (NovoMax - NovoMin) + NovoMin$$

### 2.2.1 Métodos estatísticos

A estatística prove uma fundação concreta para análise de dados, descrevendo-os em função da sua distribuição de probabilidade, matrizes de correlação de coeficientes, tabelas multidimensionais de freqüência etc., podendo constituir em ferramenta para detecção de *outliers*, considerando que os dados

comportam-se segundo uma função de densidade conhecida (SHAW, I. S.; SIMÕES, M. G., 2001).

Freqüentemente, assume-se que uma seqüência de dados aproxima-se de uma distribuição normal (apesar de que, na prática, esta premissa não seja adequada), utilizando-se da regra simples “três vezes o desvio padrão”, onde se determina que valores superiores à soma da média com o triplo do desvio padrão são *outliers*. Tal regra é inadequada, pois o estimador da variância é ampliado pela presença dos outliers, podendo encobrir algum, quando da sua aplicação. Face a mediana ser menos sensível à presença de outliers, pode-se adotá-la, aliado ao desvio absoluto (S), como medida da variabilidade dos dados (no lugar do desvio padrão), onde S é definido (PEARSON, R. K., 2002):

$$S = 1,4826 \text{ mediana } \{|X_k - X_M|\},$$

$X_k$  : dado da seqüência e  $X_M$  : mediana da seqüência de dados

Nessa abordagem, pode-se fazer uso do Boxplot, como um meio gráfico para identificar “*outliers*”, face a sua simplicidade e sua resistência à distorção causada pelos mesmos. Seu estabelecimento utiliza o conceito de quartis e a diferença interquartil, diferença entre o primeiro( $q_1$ ) e terceiro quartil( $q_3$ ), para estabelecer os limites internos e externos, definidos conforme a tabela 1 e esquematizados na figura 1:

**Tabela 1.** Limites dos *outliers* e *outliners*

Limites Internos		Limites Externos	
$f_1$	$f_3$	$F_1$	$F_3$
$q_1 - 1,5(q_3 - q_1)$	$q_1 + 1,5(q_3 - q_1)$	$q_1 - 3(q_3 - q_1)$	$q_1 + 3(q_3 - q_1)$

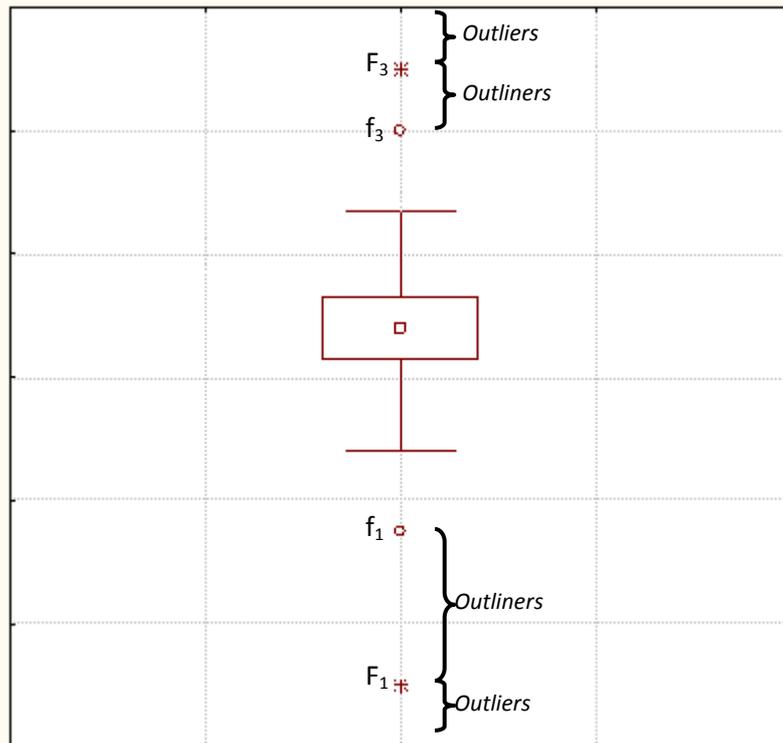


Figura 1. Limites dos outliners e outliers

Os valores concentrados além do limite  $F_3$  ou aquém do limite  $F_1$  serão considerados “outliers”, aqueles distribuídos entre os limites  $f_1$  e  $F_1$  ou  $f_3$  e  $F_3$ , “outliners”.

Entretanto há variedade de formas de determinação de “outliers”. O estabelecimento das constantes 1,5 e 3 acima demonstra o quanto uma observação pode ser divergente ou extrema, sem considerar a probabilidade destas observações serem realmente “outliers”. Assim, SCHWERTMAN, N.C. *et al.* (2004) propõem utilizar o limite semi-interquartil, baseando-se na premissa de que na prática encontram-se dados que se aproximam de uma curva normal com algum grau de assimetria, definindo os limites :

- Limite inferior ( $F_1$ )

$$F_1 = q_2 - \frac{2(q_2 - q_1)}{K_n} Z_{\alpha/2}$$

- Limite Superior ( $F_3$ )

$$F_3 = q_2 + \frac{2(q_3 - q_2)}{K_n} Z_{\alpha/2}$$

onde  $q_2$  é a mediana dos dados,  $Z_{\alpha/2}$  é valor obtido pela distribuição Normal Padronizada com um nível de confiança  $(1 - \alpha)$  e o valor de  $K_n$  é estabelecido conforme o número de valores da amostra ( $n$ ) exposto na tabela 2 abaixo:

**Tabela 2.** Valor de K (Schwertman, N.C. et al. ,2004)

n	kn	n	kn								
5	1,65798	22	1,33333	39	1,38071	56	1,34361	73	1,36635	90	1,34535
6	1,28351	23	1,40230	40	1,34165	57	1,37130	74	1,34454	91	1,36267
7	1,51475	24	1,33753	41	1,38021	58	1,34329	75	1,36557	92	1,34562
8	1,32505	25	1,40096	42	1,34104	59	1,37004	76	1,34495	93	1,36258
9	1,50427	26	1,33587	43	1,37779	60	1,34394	77	1,36543	94	1,34550
10	1,31212	27	1,39455	44	1,34226	61	1,36981	78	1,34478	95	1,36210
11	1,45768	28	1,33894	45	1,37737	62	1,34366	79	1,36474	96	1,34576
12	1,32968	29	1,39355	46	1,34175	63	1,36871	80	1,34514	97	1,36201
13	1,45268	30	1,33770	47	1,37536	64	1,34424	81	1,36461	98	1,34565
14	1,32353	31	1,38876	48	1,34278	65	1,36851	82	1,34499	99	1,36157
15	1,42975	32	1,34004	49	1,37501	66	1,34399	83	1,36398	100	1,34588
16	1,33318	33	1,38799	50	1,34235	67	1,36754	84	1,34532	200	1,34740
17	1,42684	34	1,33909	51	1,37331	68	1,34450	85	1,36387	300	1,34792
18	1,32959	35	1,38428	52	1,34322	69	1,36737	86	1,34517	400	1,34818
19	1,41322	36	1,34092	53	1,37301	70	1,34429	87	1,36330	Acima	1,34898
20	1,33568	37	1,38367	54	1,34285	71	1,36650	88	1,34548		
21	1,41132	38	1,34017	55	1,37156	72	1,34474	89	1,36319		

Para ilustrar, considere-se os valores da tabela 3 e as análises abaixo:

**Tabela 3.** Valores de exemplo ilustrativo (Schwertman, N.C. et al. ,2004)

Item	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Valor	0,534	0,535	0,57	0,45	0,548	0,431	0,481	0,423	0,475	0,486
Item	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.
Valor	0,554	0,519	0,492	0,517	0,502	0,508	0,52	0,506	0,401	0,568

- Método “três vezes o desvio padrão”, dado a média 0,501 e o desvio padrão 0,0473:

$$\text{Limite superior} = 0,501 + 3 \cdot 0,0473 = 0,6429$$

$$\text{Limite inferior} = 0,501 - 3 \cdot 0,0473 = 0,3591$$

Vê-se que não há ocorrência de dados divergentes.

- Método desvio mediano absoluto, verificando a diferença entre cada observação e a mediana tem-se os valores expostos na tabela 4:

**Tabela 4.** Valores do desvio mediano absoluto

$X_k$	0,401	0,423	0,431	0,45	0,475	0,481	0,486	0,492	0,502	0,506
$ X_k - M $	0,106	0,084	0,076	0,057	0,032	0,026	0,021	0,015	0,005	0,001
Valor	0,508	0,517	0,519	0,52	0,534	0,535	0,548	0,554	0,568	0,57
$ X_k - M $	0,001	0,01	0,012	0,013	0,027	0,028	0,041	0,047	0,061	0,063

Assim,

$$\text{Mediana} = \frac{0,508 + 0,506}{2} = 0,507$$

$$S = 1,4826 \cdot \text{Mediana}\{|X_k - \text{Mediana}|\}$$

$$S = 1,4826 \cdot \text{Mediana}\{|0,401 - 0,507|; |0,423 - 0,507|; \dots; |0,57 - 0,507|\}$$

$$S = 1,4826 \cdot \text{Mediana}\{0,106; 0,084; 0,076; \dots; 0,063\}$$

$$S = 1,4826 \cdot \frac{0,001 + 0,001}{2}$$

$$S = 1,4826 \cdot 0,001 = 0,001483$$

Portanto,

$$\text{Limite superior} = 0,507 + 3 \cdot 0,001483 = 0,511448$$

$$\text{Limite inferior} = 0,507 - 3 \cdot 0,001483 = 0,502522$$

Expondo todos os valores como *outliers*, exceto as observações 15, 16 e 18

- Por meio do gráfico Boxplot, tem-se:
  - 1º Quartil : 0,478
  - 2º Quartil (mediana) : 0,507
  - 3º Quartil : 0,5345

Utilizando-se método tradicional, não se identifica nenhum *outlier*, pois nenhum valor extrapola os limites externos, conforme exposto na tabela 5:

**Tabela 5.** Limites dos *Outliers*

Limites Internos		Limites Externos	
$f_1$	$f_3$	$F_1$	$F_3$
0,39325	0,61925	0,3085	0,704

Usando-se o método dos limites semi-interquartis, obter-se-á, para uma probabilidade de existência de um *outlier* com intervalo de confiança de 95%:

$$F_1 = 0,507 - \frac{2(0,507 - 0,478)}{1,33568} \cdot 1,645 = 0,436$$

$$F_3 = 0,507 + \frac{2 \cdot (0,5345 - 0,507)}{1,33568} \cdot 1,645 = 0,575$$

Vê-se que há valores inferiores ao limite  $F_1$ : os itens 6, 8 e 19, que não foram elencados pelo método usual.

## 2.2.2 Métodos baseados em distâncias

### I. Agrupamento

Algoritmos de classificação e agrupamento compõem o quadro deste paradigma. Agrupamento é o processo de juntar objetos similares em classes ou conjuntos (grupos, *clusters*), que são disjuntos entre si. A classificação busca, por meio de treinamento e teste com amostra conhecida, estabelecer um meio de classificar novos objetos.

Em qualquer dos casos, a representação métrica mais popular no estabelecimento da dissimilaridade/similaridade entre grupos é o cálculo da distância entre pontos, em suas mais variadas formas, como a fórmula Euclidiana, distância de Mahalonobis, Manhattan ou Mikowski (CHEN, X.; AHMAD, I. S., 2007).

Nas técnicas de agrupamento, há duas categorias: hierárquica e de particionamento. As técnicas hierárquicas são capazes de estabelecer estruturas e dividi-las recursivamente em subestruturas que representam os diversos grupos dos dados, cuja estrutura final é denominada dendograma. As técnicas de particionamento visam obter partições simples sem subdivisões e normalmente são baseadas na otimização de uma função objetivo, resultando na criação de hiper-superfícies que separam os grupos (Filippone, M. *et al.*; 2008) ( SHEN, H. *et al.*, 2006). A função

objetivo é um critério matemático que quantifica a aderência do protótipo e da partição e serve como uma função de custo que deve ser minimizada para obter a solução ótima de agrupamento via a execução de um algoritmo que estabelece a melhor decomposição do conjunto de dados em um número pré-estabelecido de grupos (OLIVEIRA, J.V.; PEDRYCZ, W., 2007) ( HE, J. *et al.*, 2004).

As técnicas de particionamento são baseadas em protótipos, que demonstram a estrutura (distribuição) de cada *cluster*. Cada protótipo é constituído por  $n$  tuplas que contém centro do cluster  $C_n$ .

Conforme OLIVEIRA, J.V.; PEDRYCZ, W. ( 2007), o algoritmo mais antigo de agrupamento é conhecido como *K-means*, ou *hard C-Means*, o qual particiona os dados em um número de grupos determinado pelo analista, baseado em um protótipo inicial também fornecido pelo analista, que deverá, por meio de algum critério, decidir se o agrupamento foi aceitável. Neste modelo, cada ponto  $x_j$  de um conjunto  $X = \{x_1, \dots, x_n\}$  é assinalado em apenas um grupo, sub-conjuntos de  $X$ . A partição em subconjuntos dir-se-á ótima quando a soma do quadrado das distâncias entre o ponto e centro do grupo para o qual foi assinalado chegar a um valor mínimo. Assim, pode-se estabelecer uma função objetivo:

$$J_h(X, U_h, C) = \sum_{i=1}^c \sum_{j=1}^n u_{ij} d_{ij}^2$$

onde,

- $C = \{c_1, \dots, c_n\}$ , o conjunto de protótipos dos grupos,
- $d_{ij}$  é a distância entre o ponto  $x_j$  e o centro do cluster  $c_i$ ,
- $U$  é a matriz de partição ( $c \times n$ ), com cada elemento  $u_{ij}$  assume o valor 0 ou 1, indicando se pertence ou não ao grupo:

$$u_{ij} = \begin{cases} 1 & \Rightarrow x_j \in c_i \\ 0 & \Rightarrow x_j \notin c_i \end{cases}$$

Estabelecendo-se as seguintes restrições:

- Cada ponto tem de pertencer a apenas um grupo:

$$\sum_{i=1}^c u_{ij} = 1, \quad \forall j \in \{1, \dots, n\}$$

- Não pode haver ponto sem pertencer a um grupo

$$\sum_{i=1}^c u_{ij} > 0, \quad \forall j \in \{1, \dots, c\}$$

A função objetivo, então, pode ser otimizada por intermédio da otimização da matriz U, mantendo-se fixa a matriz C, seguida da otimização da matriz C, fixando-se a matriz U, conforme as expressões abaixo:

- Para matriz de partição – o elemento  $x_i$  pertence ao *cluster* cuja distância seja a menor,

$$u_{ij} = \begin{cases} 1 & \text{se } i = \min_{l=1}^c (d_{lj}) \\ 0 & \text{caso contrário} \end{cases}$$

- Para matriz de protótipos - média dos vetores pertencentes a cada um,

$$C_i = \frac{\sum_{j=1}^n u_{ij} x_j}{\sum_{j=1}^n u_{ij}}$$

Este algoritmo segue as seguintes etapas:

Atribuir um valor para  $n$  (número de *clusters*)

Inicializar<sup>1</sup> a matriz de coordenadas dos centros dos *clusters*

Repetir

Para cada  $x_i$

Para cada  $c_j$

Determinar distância entre  $x_i$  e  $c_j$

Atualizar matriz  $U_{ij}$

Determinar os novos centros dos grupos  $C_i$

até que não se observe mudança em C e U.

---

<sup>1</sup> Isto pode ser feito aleatoriamente entre os vetores de dados ou estabelecendo vetores que estejam dentro da superfície que abrigue todos os dados do grupo.

O problema de obter os parâmetros que minimizem a função objetivo faz uso massivo do cálculo de distâncias entre os elementos de dados, visando buscar um centróide e estabelecer os limites que formam o agrupamento ao redor do centróide. Apresentam um custo exponencial em termos computacionais, quando se aumenta a dimensão da base de dados, o que reforça a necessidade de se reduzir a dimensão para tornar factível a operação. Paralelamente, há a tendência de este algoritmo fixar-se em um mínimo local, requerendo sua re-execução por diversas vezes com diferentes inicializações até se obter um valor melhor para a função objetivo (CHAVES, E., 2001) (AL HASAN, M., 2009).

Uma característica básica nesta abordagem é a partição rígida dos dados em agrupamentos, onde os dados obrigatoriamente pertencem a apenas um *cluster*. Esta rigidez, apesar de correta, não é adequada em todos os casos. Tal pode ser observado no exemplo abaixo (Figura 2), onde dois agrupamentos circulares apresentam interseção de dados, que serão atribuídos a um cluster, apesar de distar igualmente dos dois clusters. Tal situação é inadequada, pois este dado pertence aos dois agrupamentos (HÖPPNER, F. *et al.*, 1999) (ZHANG, D.; CHEN, S., 2004).

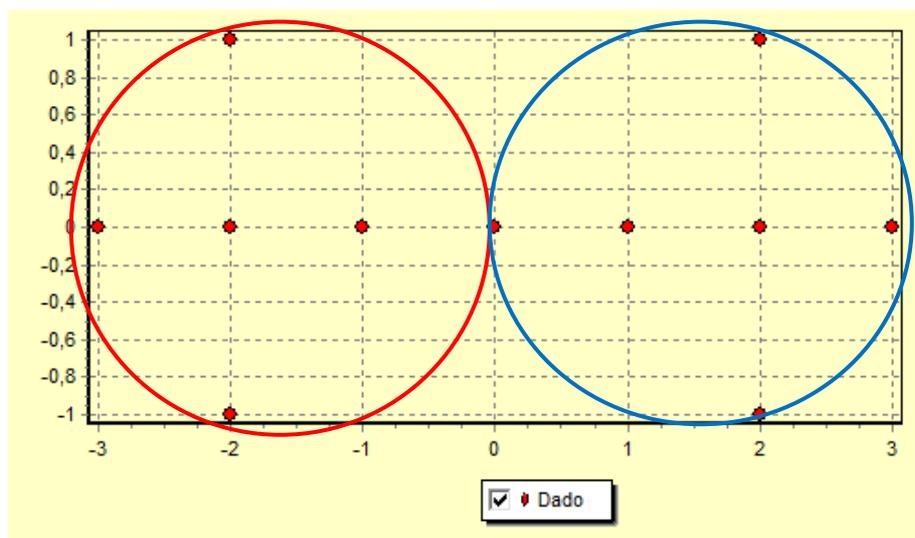


Figura 2 Agrupamento Rígido

Desta forma, a solução deste problema pode ser obtida pela introdução de graus de pertinência a cada um dos agrupamentos, baseado nos conceitos de conjunto nebulosos (*Fuzzy sets*) (OLIVEIRA, J.V.; PEDRYCZ, W., 2007). Nesta abordagem, permite-se o relacionamento dos dados a todos os *clusters* de acordo com graus de pertinência contido no intervalo  $[0,1]$ , o que garante a flexibilidade de um

dado pertencer a mais de um *cluster*, ou seja, as fronteiras entre os grupos não são rigidamente estabelecidas (XU, R.; WUNSCH, D., 2005) (DÖRING, C. *et al.*, 2006).

Assim, considerando um conjunto de dados  $X = \{x_1, x_2, x_3, \dots, x_n\}$  e conjunto de partições  $C = \{c_1, c_2, c_3, \dots, c_c\}$ , cada dado  $x_j$  será associado a um *cluster*  $c_i$  com um grau de pertinência  $u_{i,j}$  contido no intervalo  $[0,1]$ . Este indicador não constitui mais um marcador do agrupamento a qual o dado pertence, mas um vetor dos graus de pertinência do dado a cada cluster. O conjunto dos vetores de pertinência constituir-se-á, portanto, na matriz de partição fuzzy, que retratará a ambigüidade dos agrupamentos sobrepostos.

O grau de pertinência permitido variará de acordo com a restrição que se imponha e a forma de interpretação deste. Este algoritmo é também denominado agrupamento probabilístico (*Probabilistic Fuzzy C-means – FCM*), uma vez que o grau de pertinência de um determinado dado  $x_i$  é formalmente a probabilidade deste dado pertencer a um determinado *cluster*  $c_c$  (TIMM, H. *et al.*, 2004).

Para tal, as seguintes restrições são impostas :

- (1)  $\sum_{j=1}^n u_{ij} > 0 \quad \forall i \in \{1, \dots, c\}$  e
- (2)  $\sum_{i=1}^c u_{ij} = 1 \quad \forall j \in \{1, \dots, n\}$  .

A primeira restrição garante que não haverá agrupamento vazio e a segunda impõe que a soma de todos os graus de pertinências seja igual a um, significando que os graus de pertença de cada dado são distribuídos por todos os *clusters*.

Neste caso, busca-se estabelecer os valores que minimizem a função objetivo:

$$J(X, U, C) = \sum_{i=1}^c \sum_{j=1}^n U_{ij}^m d_{ij}^2$$

O parâmetro  $m$ ,  $m > 1$ , é denominado fuzzyficador ou peso, cujo objetivo é flexibilizar os limites das partições. Para valores altos, as fronteiras ficam suavizadas e mais rígidas para valores baixos. Para valores próximos de um, tem-se valores obtidos pelo algoritmo K-means. Usualmente, o valor deste parâmetro é instado em 2, sendo o valor um inviável pois levaria a uma divisão por zero quando se buscasse determinar a matriz de partição (FILIPPONE, M. *et al.*, 2007) (OLIVEIRA, J.V.; PEDRYCZ, W. , 2007).

Conforme expõe TIMM, H. *et al.* (2004), a função de otimização não pode ser minimizada diretamente. Usa-se, então, um algoritmo que otimiza alternativamente a matriz de pertinência e matriz dos centros dos *clusters*. Assim, inicia-se pela otimização da matriz de pertinência, fixando-se o conjunto dos agrupamentos e, em seguida, otimiza-se a matriz dos centros dos agrupamentos, fixando-se a matriz de pertinência, tomando-se as equações abaixo:

- Matriz de partição (pertinência):

$$U_{ij} = \frac{d_{ij}^{\frac{2}{m-1}}}{\sum_{l=1}^c d_{ij}^{\frac{2}{m-1}}}$$

Esta equação demonstra claramente o comportamento probabilístico do grau de pertinência, pois não depende unicamente da distância do dado ao cluster, mas de todas as distâncias deste dado a todos os clusters.

- Matriz de clusters:

$$C_i = \frac{\sum_{j=1}^n U_{ij}^m x_j}{\sum_{j=1}^n U_{ij}^m}$$

Assim, o algoritmo de agrupamento, segundo essa abordagem, segue as seguintes etapas:

Estabelecer o número de clusters, o número de iterações máximo, o valor de  $m$  e valor  $\epsilon$

Inicializar matriz de protótipos  $C$  (centro do cluster)

Computar  $E = 0$

Repetir Enquanto o

    Atualizar matriz de pertinência ( $U_{ij}$ )

    Atualizar matriz de centro de cluster ( $C_i$ )

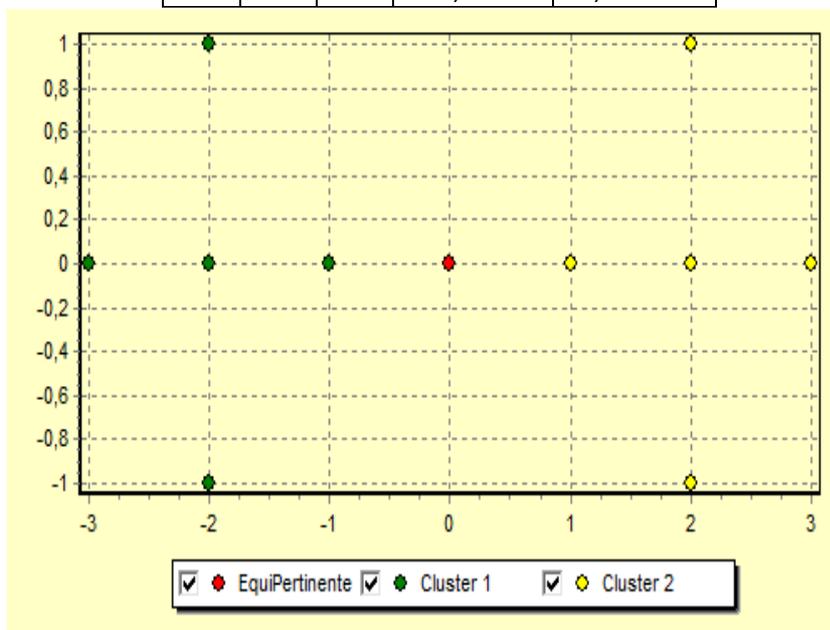
    Computar  $E =$  valor absoluto da maior diferença entre a matriz de pertinência calculada anteriormente e a atual

Até que número da iteração  $>$  número máximo de iterações ou o valor de  $E < \epsilon$

O algoritmo FCM pode, desta forma, expressar diferentemente o grau de pertinência de cada ponto, onde cada um tem um valor proporcional à similaridade com os pontos centrais de cada cluster. Diferentemente de partições rígidas, onde não se estabelecem diferenças entre pontos próximos ou distantes do centro do cluster, aqui, os valores de pertinência indicam tal situação, provendo valores altos para pontos próximos do centro e baixo para os distantes (observe-se tabela 6 e figura 3 abaixo):

**Tabela 6.** Matriz de pertinência

j	Pontos		Grau de pertinência	
	x1	x2	Cluster 1	Cluster 2
1	-3	0	0,99999937	6,26E-07
2	-2	0	1	1,59E-14
3	-1	0	0,99999488	5,12E-06
4	-2	1	0,99999883	1,17E-06
5	-2	-1	0,99999883	1,17E-06
6	0	0	0,5	0,5
7	1	0	5,12E-06	0,9999949
8	2	0	1,59E-14	1
9	3	0	6,26E-07	0,9999994
10	2	1	1,17E-06	0,9999988
11	2	-1	1,17E-06	0,9999988



**Figura 3** Agrupamento FCM

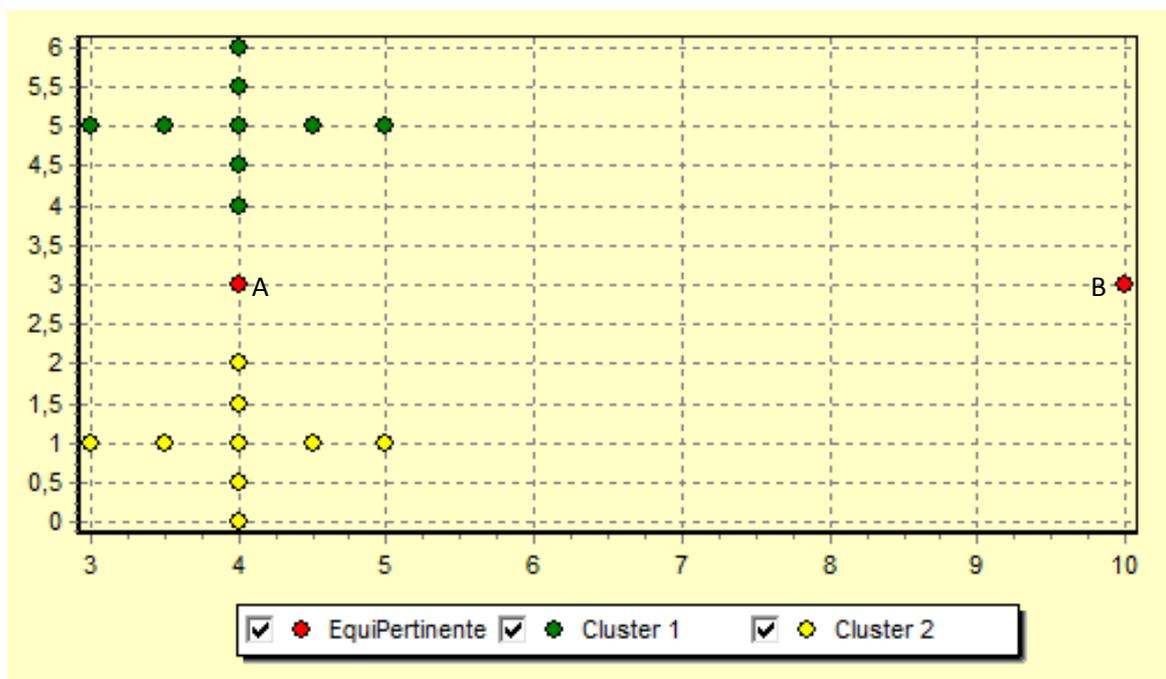
Apesar de desejável, a pertinência relativa a vários *clusters* pode levar a caracterizações indevidas. Valores baixos de pertinência a todos *clusters* podem, à primeira vista, ser entendidos como algo normal, quando pode revelar, em verdade,

um dado divergente (OLIVEIRA, J.V.; PEDRYCZ, W., 2007). Dados eqüidistantes do centro dos *clusters* podem apresentar graus de pertinência diferentes, posto que o grau de pertinência de um ponto em um cluster é um número relativo, que depende do grau de pertinência deste ponto aos outros clusters (KRISHNAPURAN, R.; KELLSER, J.M., 1993) (KRISHNAPURAN, R.; KELLSER, J.M., 1996).

Assim, tome-se os dados da tabela 7, que submetidos ao algoritmo FCM com dois clusters, apresenta o agrupamento exposto na figura 4, cujos graus de pertinência são apresentados também na tabela 7:

**Tabela 7.** Coordenadas cartesianas dos pontos e graus de pertinência

j	Pontos		Grau de pertinência		j	Pontos		Grau de pertinência	
	x	y	Cluster 1	Cluster 2		x	y	Cluster 1	Cluster 2
1	4	5	1	1,55E-11	11	10	3	0,5	0,5
2	3,5	5	1	1,09E-07	12	4	1	1,55E-11	1
3	3	5	0,99999	1,03E-05	13	3,5	1	1,09E-07	1
4	4,5	5	1	7,13E-12	14	3	1	1,03E-05	0,99999
5	5	5	1	7,64E-08	15	4,5	1	7,13E-12	1
6	4	5,5	1	1,76E-08	16	5	1	7,64E-08	1
7	4	6	0,999999	9,39E-07	17	4	1,5	9,59E-10	1
8	4	4,5	1	9,59E-10	18	4	2	6,97E-06	0,999993
9	4	4	0,999993	6,97E-06	19	4	0,5	1,76E-08	1
10	4	3	0,5	0,5	20	4	0	9,39E-07	0,999999



**Figura 4** Algoritmo FCM – Registros equipertinentes

Intuitivamente, considerando os pontos A e B, vê-se que são pontos não pertencentes a nenhum dos agrupamentos. O ponto A, mais próximo dos centros dos clusters,

deveria ter um alto grau de pertinência e o ponto B, distante, um baixo grau (pensando-se em grau de pertencimento ou tipicidade). O algoritmo, entretanto, assinala ambos com valor 0,5, que não representa a pertença do ponto ao cluster, nem pode ser entendido com o grau de semelhança, pouco ou muito semelhante.

Isto é conseqüência da restrição de que a soma de todas as pertinências deve ser igual a um, ou seja, há que se distribuir este valor por todos os clusters (vide (2) acima), nem sempre representando corretamente o grau de pertinência ao agrupamento (XU, R.; WUNSCH, D., 2005)( XIE, Z., *et al.*, 2008).

Para evitar este problema deve-se eliminar tal restrição, definindo-se um novo algoritmo de agrupamento, *Possibilistic Fuzzy C-means* (PCM), assim denominado em virtude do grau de pertinência inferir uma possibilidade do dado pertencer a um determinado cluster. Para tanto, onde as restrições anteriores são limitadas a

$$\sum_{i=1}^c u_{ij} > 0 \quad \forall j \in \{1, \dots, c\} .$$

Esta restrição determina uma alteração na função objetivo, pois, caso contrário, pode se obter uma solução inviável. A minimização da função objetivo pode levar a  $U_{ij} = 0$  para todo  $i \in \{1, \dots, c\}$  e  $j \in \{1, \dots, n\}$ , ou seja, nenhum dado é assinalado à cluster algum e todos os clusters são vazios. Para evitar tal possibilidade, insere-se um peso à equação para forçar que os graus de pertinência se afastem do valor nulo, modificando-se a função objetivo para:

$$J(X, U, C) = \sum_{i=1}^c \sum_{j=1}^N U_{ij}^m d_{ij}^2 + \sum_{i=1}^c \eta_i \sum_{j=1}^n (1 - U_{ij})^m$$

onde  $\eta_i > 0$ , ( $i = 1, \dots, c$ ).

Assim, enquanto o primeiro termo da equação busca minimizar as distâncias, a segunda, evitar a solução indesejada, penalizando os dados que se afastam do centro (OLIVEIRA, J.V.; PEDRYCZ, W., 2007) (CHINTALAPUDI, K. K.; KAM, M., 1998) (KRISHNAPURAN, R; KELLSER, J.M., 1993).

De igual forma ao algoritmo anterior, deriva-se da função objetivo as equações implementadas no algoritmo:

- Matriz de partição (pertinência):

$$U_{ij} = \frac{1}{1 + \left(\frac{d_{ij}^2}{\eta_i}\right)^{\frac{1}{m-1}}}$$

Vê-se que a pertinência depende da distância do dado  $X_j$  ao centro do cluster  $C_i$ , pequenas distâncias, alto grau de pertinência e vice-versa.

- Matriz de clusters:

$$C_i = \frac{\sum_{j=1}^n U_{ij}^m x_j}{\sum_{j=1}^n U_{ij}^m}$$

Note-se que esta formulação é idêntica àquela instada no algoritmo FCM.

- Peso:

$$\eta_i = \frac{\sum_{j=1}^n U_{ij}^m d_{ij}^2}{\sum_{j=1}^n U_{ij}^m}$$

Assim, o algoritmo de agrupamento, segundo essa abordagem, segue as seguintes etapas (HÖPPNER, F. *et al.*, 1999):

Estabelecer o número de clusters, o número de iterações máximo, o valor de  $m$  e valor  $\varepsilon$

Executar o algoritmo FCM

Inicializar matriz de protótipos (centro do cluster)

Repetir 2 vezes

Computar  $E = 0$

Estimar o valor de  $\eta_i$

Repetir

Atualizar matriz de pertinência ( $U_{ij}$ )

Atualizar matriz de centro de cluster ( $C_i$ )

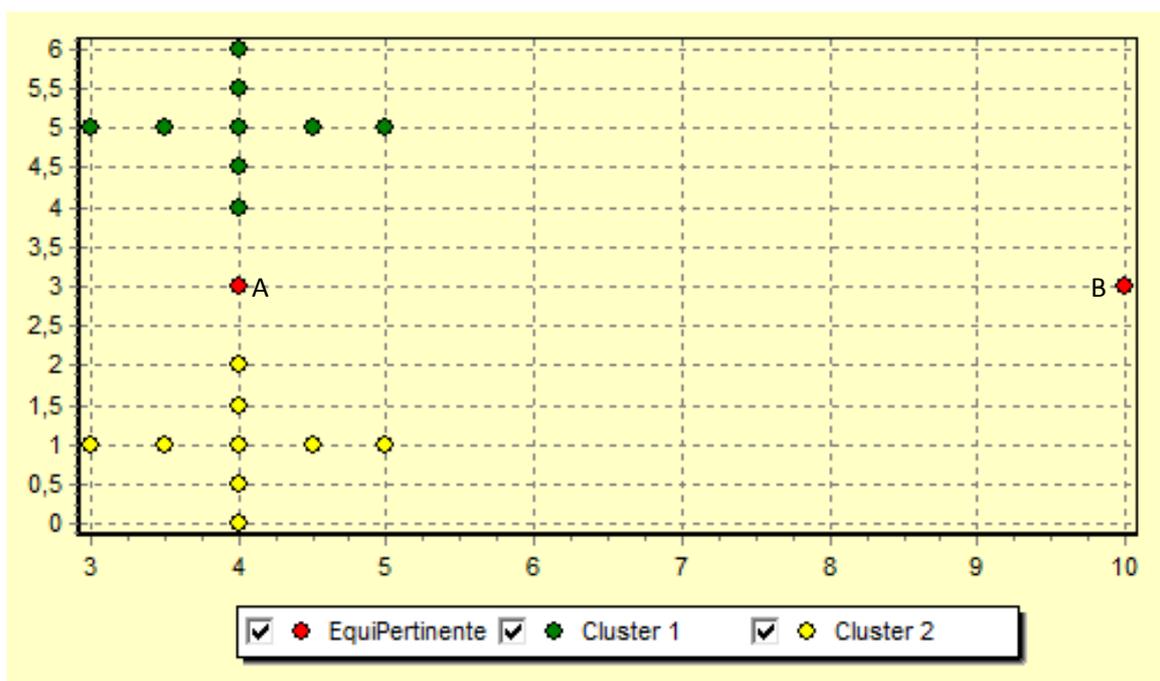
Computar  $E$  = valor absoluto da maior diferença entre a matriz de pertinência calculada anteriormente e a atual

Até que número da iteração > número máximo de iterações ou o valor de  $E < \varepsilon$

Tomando-se, novamente, os dados expostos na tabela 8 e aplicando-os a esse algoritmo, estabelecem-se os seguintes graus de pertinência colocados na mesma tabela e agrupamento observado no figura 5:

**Tabela 8.** Coordenadas cartesianas dos pontos e graus de pertinência (PCM)

j	Pontos		Grau de pertinência		j	Pontos		Grau de pertinência	
	x	y	Cluster 1	Cluster 2		x	y	Cluster 1	Cluster 2
1	4	5	1	5,63E-08	11	10	3	5,76E-10	5,76E-10
2	3,5	5	0,983723	5,21E-08	12	4	1	5,63E-08	1
3	3	5	0,05573	4,16E-08	13	3,5	1	5,21E-08	0,983723
4	4,5	5	0,983723	5,21E-08	14	3	1	4,16E-08	0,05573
5	5	5	0,05573	4,16E-08	15	4,5	1	5,21E-08	0,983723
6	4	5,5	0,983721	1,73E-08	16	5	1	4,16E-08	0,05573
7	4	6	0,055728	6,04E-09	17	4	1,5	2,14E-07	0,983724
8	4	4,5	0,983724	2,14E-07	18	4	2	1,00E-06	0,055732
9	4	4	0,055732	1,00E-06	19	4	0,5	1,73E-08	0,983721
10	4	3	5,76E-05	5,76E-05	20	4	0	6,04E-09	0,055728



**Figura 5** - Algoritmo PCM - Evidência de outlier

Aparentemente, não houve mudanças, entretanto, pode-se verificar que os pontos A e B receberam um grau de pertinência pequeno, especialmente B.

Esta é a diferença marcante entre os dois métodos: enquanto o algoritmo FCM busca imputar o dado a um cluster, o PCM não o faz! Desta forma, o algoritmo PCM tem a capacidade de interpretar dados como *outliers*, provendo um baixo grau de

pertinência a esses dados em relação a todos os clusters (valores próximos a zero) (OLIVEIRA, J.V.; PEDRYCZ, W., 2007).

### 2.2.3 – Métodos baseados em similaridade

Aplicando-se o algoritmo PCM a um conjunto de dados referentes a lâmpadas de veículos obtido em <http://fuzzy.cs.unimagdeburg.de/clusterbook>, sabendo-se que há quatro clusters conforme descrito em HÖPPNER, F. *et al.* (1999), obtém-se o resultado exposto na figura 6 abaixo:

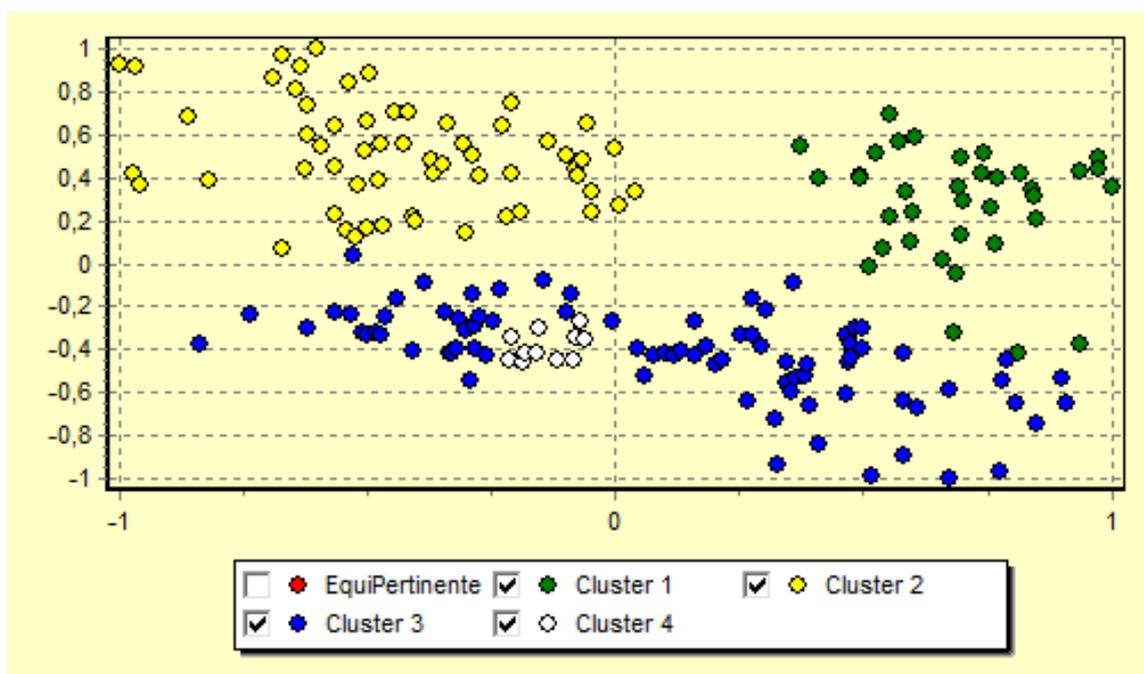


Figura 6 Agrupamento de lâmpadas de veículos - Algoritmo PCM

Nota-se uma confusão no estabelecer dos agrupamentos, sendo notória a falha do algoritmo, que se deve ao uso da distância euclidiana como medida de similaridade.

Diversas variações têm sido propostas para substituí-la, uma destas, o uso de funções kernel, o que significa mapear os dados do espaço original  $X$  para um espaço de maior dimensão  $H$  (espaço característico),  $\phi: X \rightarrow H$ , por meio desta função. Tal transformação possibilita a aplicação de um classificador linear no espaço kernel que resulta em uma classificação não linear no espaço original. (PETROVSKIY, M. I., 2003) (PÉREZ-CRUZ, F.; BOUSQUET, O., 2004) (CAMASTRA, F.; VERRI, A., 2005) (GRAVES, D.; PEDRYCZ, W., 2007) (ZHANG, H. *et al.*, 2007).

Entretanto, não se requer transformar os dados e trabalhar no espaço kernel, mas utilizar uma forma de obter tal produto neste espaço multidimensional sem

mapear diretamente os dados para este espaço, ou seja,  $K(x, x') = \langle \phi(x), \phi(x') \rangle$  (SCHÖLKOPF, B; SMOLA, A. J., 2002).

Neste mister, considere-se o problema de discriminação conforme visto na figura 7, onde se deseja estabelecer uma divisão entre os dois grupos, que é claramente visto como uma fronteira não linear.

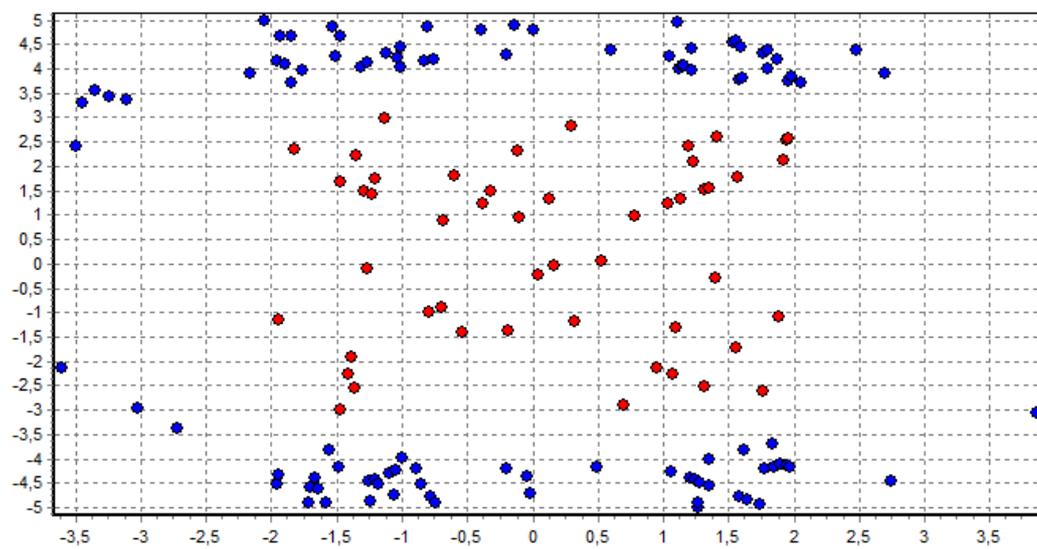


Figura 7 Dados no espaço original

Aplica-se, então, uma função de mapeamento  $\phi(x, y) = (x^2, \sqrt{2}xy, y^2)$  e se verifica que a fronteira entre dois grupos torna-se linear, conforme exposto na figura 8 abaixo:

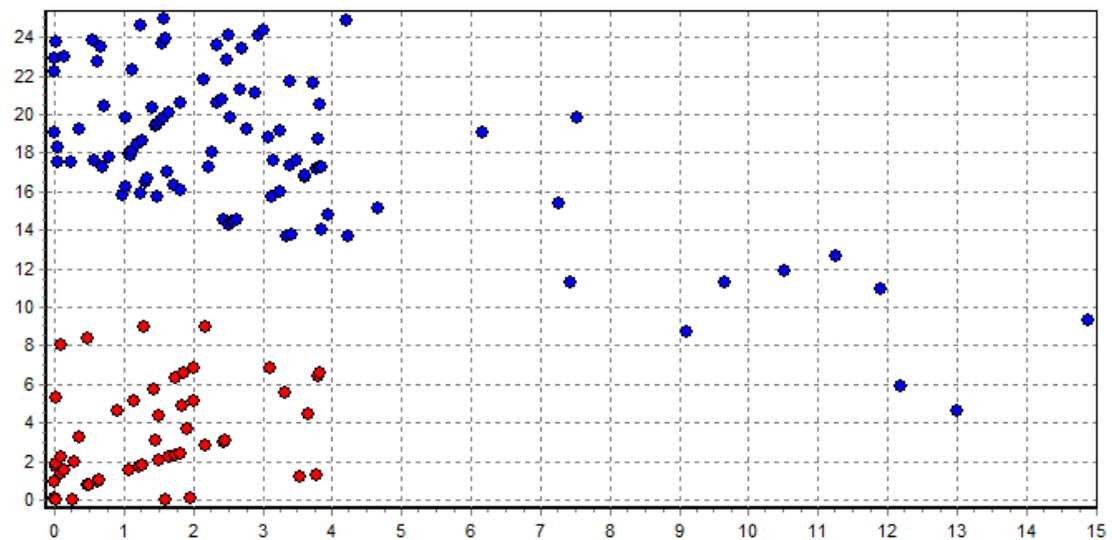


Figura 8 Projeção dos dados no espaço kernel

Aí, simplesmente, houve uma transformação dos dados esperando-se que uma estrutura linear emergisse, o que, algumas vezes, demanda buscar maiores dimensões, tornando a tarefa um pouco mais complicada, levando a se restringir aos espaços cobertos pelo produto interno (PÉREZ-CRUZ, F.; BOUSQUET, O., 2004). Tal fato possibilita formular o *kernel trick*, onde se pode estabelecer o produto interno a partir de  $x$  e  $x'$ , sem explicitamente aplicar-se os operadores  $\phi(x)$  e  $\phi(x')$ .

Conforme SCHÖLKOPF, B; SMOLA, A. J. (2002), PÉREZ-CRUZ, F.; BOUSQUET, O. (2004) e WU, K.; WANG, S. (2009), há as seguintes funções kernel mais utilizadas:

- Linear :  $K(x, x') = \langle \phi(x), \phi(x') \rangle$
- Polinomial de grau  $d$ :  $K(x, x') = (a + \gamma \langle x, x' \rangle)^d, \gamma > 0$
- Gaussiano ou RBF :  $K(x, x') = \exp\left(-\|x - x'\|^2 / 2\sigma\right)$
- Sigmóide :  $K(x, x') = \tanh(\gamma \langle x, x' \rangle + r)$

Destas escolhas, o kernel gaussiano apresenta-se como uma opção mais usual de aplicação (ZHANG, L. *et al.*, 2006).

A característica básica do uso de funções kernel é sua aplicação de forma fácil sobre dados de tipos variados e o dispensar do conhecimento da natureza dos dados.

### I. Agrupamento Nebuloso com função Kernel (KPCM)

Desta forma, introduzindo a função kernel às equações do algoritmo FCM, ter-se-á :

- Matriz de partição (pertinência):

$$U_{ij} = \frac{\left(1 - K(x_j, c_i)\right)^{-1/(m-1)}}{\sum_{i=1}^c \left(1 - K(x_j, c_i)\right)^{-1/(m-1)}}$$

- Matriz de clusters:

$$C_i = \frac{\sum_{j=1}^n U_{ij}^m K(x_j, c_i) x_j}{\sum_{j=1}^n U_{ij}^m K(x_j, c_i)}$$

De igual forma ao algoritmo PCM:

- Matriz de partição (pertinência):

$$U_{ij} = \frac{1}{1 + 2 \left( \frac{1 - K(x_j, c_i)}{\eta_i} \right)^{-1/(m-1)}}$$

- Matriz de clusters:

$$C_i = \frac{\sum_{j=1}^n U_{ij}^m K(x_j, c_i) x_j}{\sum_{j=1}^n U_{ij}^m K(x_j, c_i)}$$

- Peso:

$$\eta_i = \frac{\sum_{j=1}^n U_{ij}^m 2(1 - K(x_j, c_i))}{\sum_{j=1}^n U_{ij}^m}$$

Os algoritmos permanecem absolutamente iguais àqueles já expostos, alterando-se apenas a forma de se estabelecer a matriz de pertinência e a matriz de protótipos (centro dos *clusters*).

Aplicando-se o algoritmo PCM com função kernel a um conjunto de dados referentes às lâmpadas de veículos citados anteriormente, com quatro *clusters*, obtém-se o agrupamento exposto na figura 9:

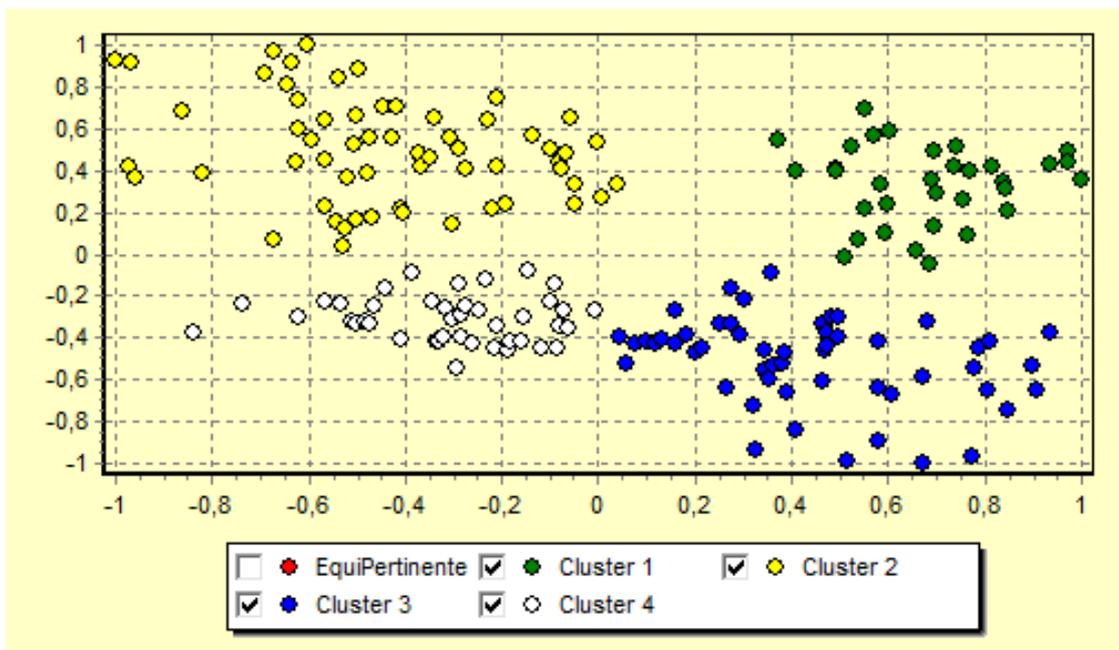


Figura 9 Algoritmo KPCM - Agrupamento de lâmpadas de veículos

Observa-se que o algoritmo KPCM logrou sucesso em separar todos os subconjuntos de dados.

## II. Máquinas de vetor suporte

O algoritmo denominado máquinas de vetor suporte (*Support Vector Machine* - SVM) é um método de aprendizado supervisionado, baseado na teoria de aprendizado estatístico, aplicado a problemas de classificação e regressão (LI, K. e TENG, G., 2006) (HOREWICZ, M. C *et al.*, 2007).

Conforme exposto por DOMÍNGUES, R. A.; NANDI, A. K. (2009), BERRUETA, L. A. *et al.* (2007), MÜLLER, K. R. *et al.* (2001) e LAM, K. *et al.* (2008), ao se considerar um problema onde duas classes são separáveis linearmente, fica fácil de se ver que a melhor função discriminante é um hiperplano que fica em alguma região entre os grupos das duas classes. Assim, este método busca obter uma fronteira entre duas classes independentemente da função de distribuição de probabilidade dos dados do conjunto.

Desta forma, observe-se a figura 10 abaixo, onde dois grupos podem ser separados por diversas fronteiras. Entre estas possibilidades, uma maximiza a margem de separação, isto é, maximiza a distância entre esta e os pontos mais próximos de cada classe. Esta margem é denominada como hiperplano ótimo (HOREWICZ, M. C. 2007), (LIN, C. *et al.*, 2008) (GUNN, S. R., 1998).

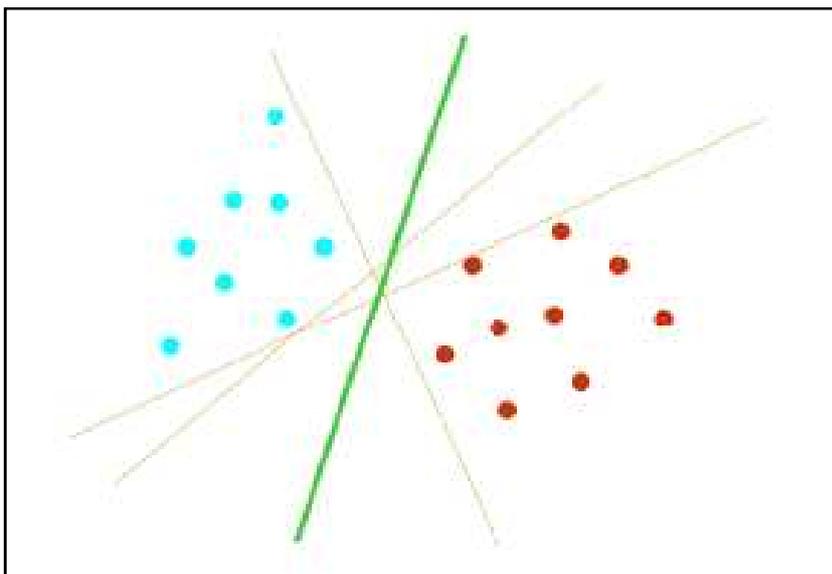
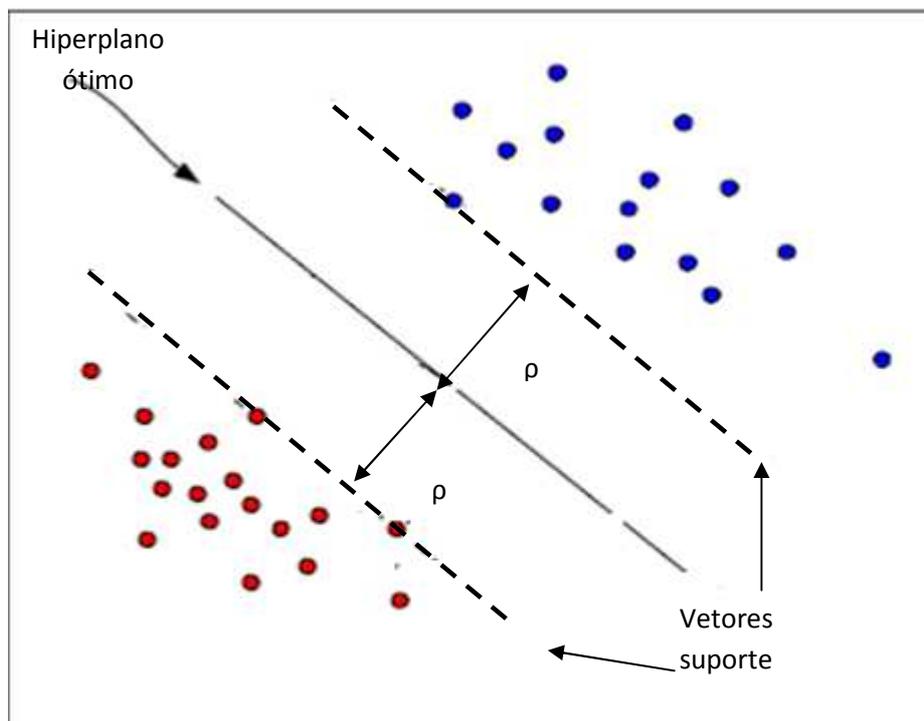


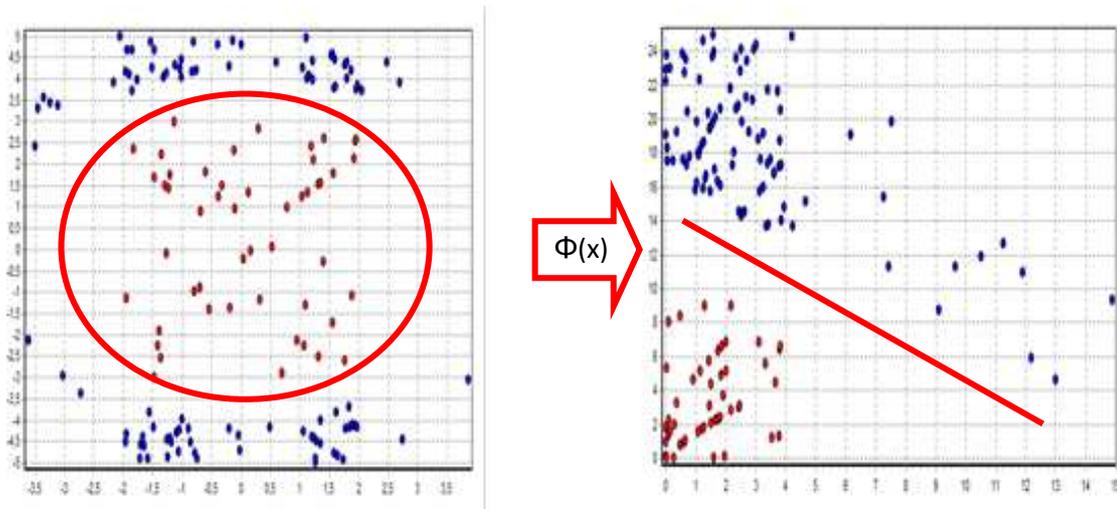
Figura 10. Hiper-plano de separação

Assim, considerando um vetor de pesos  $w \in \mathbb{R}^m$  e um bias  $b \in \mathbb{R}$ , a superfície de separação será constituída por um hiperplano na forma  $g(x) = (w^T x) + b = 0$ . Em problemas linearmente separáveis, existirão infinitos hiperplanos, dentre os quais um em particular, que maximize a margem de separação. Esta melhor região para estabelecer tal fronteira é caracterizada por um pequeno número de pontos de cada classe, que se encontram à distância  $\rho$  do hiperplano ótimo, e são denominados vetores de suporte (*support vectors*), conforme exposto na figura 11. Estes vetores exercem um papel importante, posto que são os pontos mais próximos da superfície de decisão e, desta forma, os de mais difícil classificação.



**Figura 11.** Hiperplano Ótimo

Na maioria dos casos reais os dados não são linearmente separáveis, não sendo possível separar os dados por um hiperplano. Recorre-se, então, ao mapeamento não-linear de um vetor de entrada de dimensão  $n$  em um espaço característico de dimensão  $k$ , sendo  $k > n$ , através de um mapeamento  $\phi(x)$ , uma das funções de kernel já explanadas anteriormente, onde as classes são linearmente separáveis (observe figura 12).

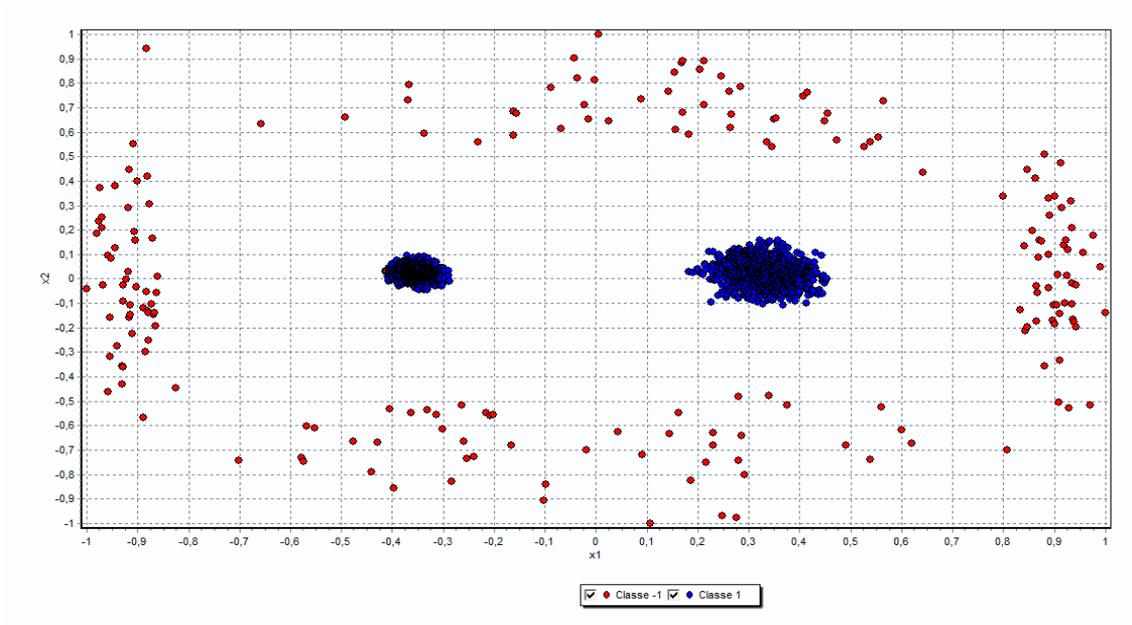


**Figura 12.** Mapeamento para o espaço Kernel

A obtenção deste produto interno no espaço característico é calculado diretamente como uma função do espaço original,  $K(x,y) = \phi(x) \cdot \phi(y)$ , conhecido como kernel *trick* (SHIN, H. J. et al., 2005).

Máquinas de vetor suporte são técnicas supervisionadas de reconhecimento de padrões, requerem um conjunto de treinamento, onde as amostras já estão categorizadas, para derivarem um modelo para identificar amostras desconhecidas. No caso do estudo aqui desenvolvido, desconhecem-se os grupos de dados, caracterizando uma tarefa não supervisionada.

Há, entretanto, uma variante de máquinas de vetor suporte, *SVM One-Class*, onde se considera que os dados normais são aderentes a uma determinada função de densidade de probabilidade e, assim, encontram-se próximos uns dos outros, enquanto os dados divergentes têm comportamento aleatório, estando distanciados dos primeiros (YANG, J. et al., 2007). Da forma semelhante, busca transformar os dados para o espaço característico de maior dimensão por meio de uma função kernel e separá-los da origem com uma margem máxima (SCHÖLKOPF, B; SMOLA, A. J., 2002). O objetivo é encontrar um hiperplano mais distante da origem, estabelecendo-se dois grupos, os mais próximos da origem, considerados normais (classe +1), e os mais distantes, anormais (classe -1), conforme esquematizado na figura 13.



**Figura 13.** Separação dos dados provida pelo algoritmo SVM one-class

Conforme exposto por LI, K. e TENG, G. (2006), WANG, Y. et al. (2004), TRAN, Q. et al. (2003), BOSE, R. P. J. C. e SRINIVASAN, S.H. (2005), GUO, Q. et al. (2005), WANG, D. et al. (2006), SCHÖLKOPF, B. e SMOLA, A. J. (2002), LIN, C. et al. (2008) e ONODA, T. et al. (2007), a estratégia, portanto, é mapear os dados  $X = \{x_1, x_2, x_3, \dots, x_m\}$  para o espaço kernel por meio de uma função  $\phi(x_i)$ , separando-os da origem com uma margem máxima, por meio da minimização da função:

$$\frac{1}{2} \|w\|^2 + \frac{1}{\nu m} \sum_{i=1}^m \varepsilon_i - \rho$$

$$\text{sujeita a } \langle w, \phi(x_i) \rangle \geq \rho - \varepsilon_i, \quad \varepsilon_i \geq 0$$

onde  $w$  é o vetor de pesos,  $\|w\|$  é o módulo do hiperplano ótimo,  $\varepsilon_i$  é a variável de folga responsável por flexibilizar a restrição de separação,  $\nu$  é um parâmetro que busca mediar dois objetivos conflitantes minimizar a folga e maximizar a margem de separação,  $\rho$  é a distância entre os vetores suporte e a margem de separação. Como os dados podem não ser linearmente separáveis no espaço original, utiliza-se uma função kernel  $\phi(x_i) = K(x_i, x'_i) = e^{-\|x_i - x'_i\|^2 / 2\sigma}$ , que possibilita tal ocorrência num espaço característico de maior dimensão.

Como a solução é provida por  $w$  e  $\rho$ , estabelece-se a seguinte função de decisão:

$$f(x) = \text{sgn}(\langle w, \phi(x_i) \rangle - \rho)$$

$$\text{onde } \langle w, \phi(x_i) \rangle - \rho \begin{cases} \geq 0, & f(x) = +1 \\ < 0, & f(x) = -1 \end{cases}$$

Assim, este método identifica os dados divergentes entre aqueles positivos, classe um ou *one-class*, e classifica como negativos, os *outliers*. Por outras palavras, a função devolverá um valor +1 para os dados contidos em região que contém a maior parte dos dados e -1 para aqueles que não estiverem nesta região (HAO, P., 2008).

### III. Função de similaridade média

Técnicas baseadas em proximidade são simples de serem implementadas e dispensam qualquer conhecimento prévio acerca dos dados. Como são baseadas no cálculo exaustivo de distâncias entre todos os registros, demandam um esforço computacional maior, conforme cresce a dimensão e o número de registros (HODGE, V.; AUSTIN, J., 2004). Tal, entretanto, não é fator impeditivo, face o estado atual dos processadores.

Neste viés considera-se como anômalo todo registro que tenha mais de  $K$  vizinhos com distâncias superiores a  $d$ , onde  $d$  e  $K$  são parâmetros informados. Apesar de computacionalmente possível e intuitivo, este procedimento carrega algumas dificuldades, a começar pelo estabelecimento da distância  $d$ , que não é trivial (KNORR, E.; NG, R., 1998)( VOULGARIS, Z.; MAGOULAS, G, 2008).

RAMASWAMY, S. *et al.*, (2000) expõem que, para a evidenciação de um *outlier*, pode se prescindir do parâmetro de distância  $d$ , baseando-se na distância do  $k$ -ésimo vizinho mais próximo de um ponto. Intuitivamente, percebe-se que a distância é uma medida de divergência. Denotando  $d(k,p)$  como a distância entre um ponto  $p$  e seu  $k$ -ésimo vizinho mais próximo, observa-se que para valores baixos de  $d(k,p)$  ter-se-á uma vizinhança densa e para valores altos, vizinhanças mais esparsas, o que caracterizaria um *outlier*. Assim, os autores informam que, geralmente, o analista está interessado nos maiores  $n$  *outliers* e, em virtude disso, estabelecem a seguinte norma: dado  $K$  (número de vizinhos) e  $n$  (número de *outliers*), um ponto  $p$  será considerado um *outlier* se não mais do que  $n-1$  outros pontos do conjunto tiverem distâncias superiores a  $d(k,p)$ . Por outras palavras, os top  $n$  pontos com os maiores  $d(k,p)$  serão considerados *outliers*.

De forma semelhante, ANGIULLI, F. e PIZZULI, C. (2005) estabelecem um grau de isolamento de um dado em relação aos seus vizinhos mais próximos como sendo a soma de todas as distâncias do ponto p aos K vizinhos mais próximos, denominando-se a este resultado como o peso do ponto p. Quanto maior o peso, maior o isolamento, sendo os pontos com maiores pesos considerados *outliers*.

Esses métodos carecem de estimar-se o valor da distância e o número de vizinhos considerados próximos.

Assim, propõe-se um método não paramétrico, que dispense qualquer conhecimento prévio. Considerando a existência de um conjunto de dados  $X = \{x_1, x_2, \dots, x_n\}$ , a cada  $x_i$  atribui-se uma medida de semelhança, que é estabelecida pelo cálculo das similaridades entre i-ésimo dado e todos os demais (n-1) pontos do conjunto por intermédio de uma função de kernel:

$$k(i, j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma}\right), i = 1..n, j = 1..n, i \neq j$$

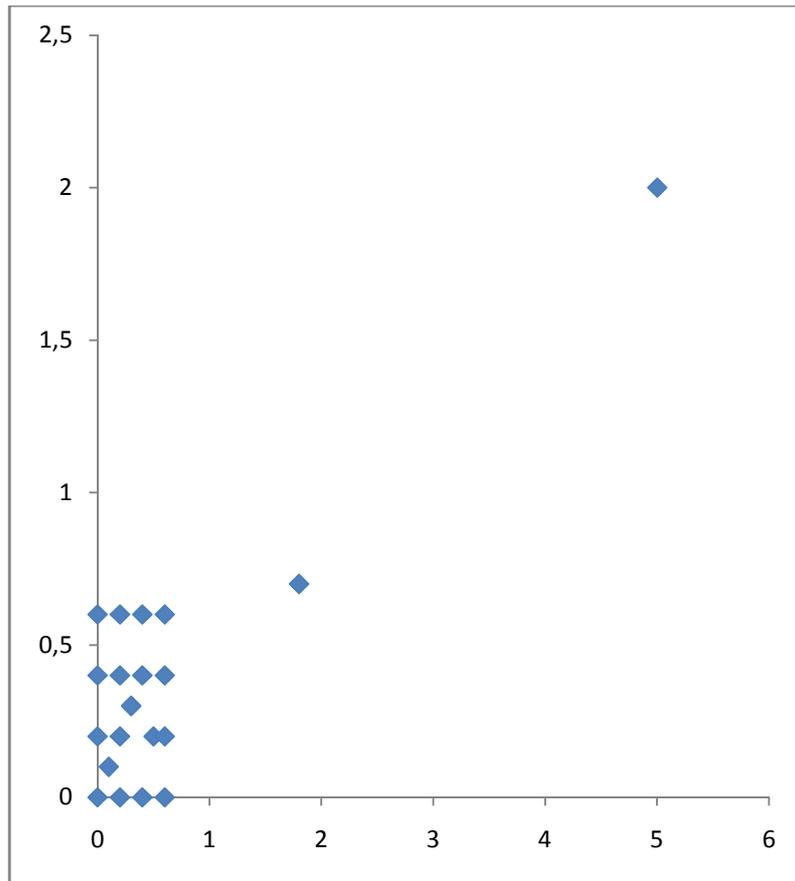
A medida de semelhança de um registro  $x_i$  será a média das similaridades entre este registro e todos os demais:

$$S_i = \frac{1}{(n-1)} \sum_{j=1}^n k(i, j), \quad j \neq i$$

Aos valores obtidos, verifica-se que os de menor monta caracterizam um dado anômalo, como pode ser observado no conjunto de valores aleatórios expostos na tabela 9 e figura 14 abaixo:

**Tabela 9.** Dados Aleatórios

n	x	y	n	x	y
1	5	2	10	0	0,6
2	1,8	0,7	11	0,2	0
3	0,5	0,2	12	0,4	0
4	0,2	0,6	13	0,3	0,3
5	0	0	14	0,6	0
6	0,4	0,6	15	0,4	0,4
7	0,2	0,2	16	0	0,4
8	0	0,2	17	0,6	0,4
9	0,6	0,6	18	0,1	0,1



**Figura 14** Dados Aleatórios

Nota-se que o ponto (1) é claramente um *outlier*, cujo índice de semelhança reforça esta conclusão, mormente se comparado com os restantes, conforme se observa na tabela 10:

**Tabela 10.** Grau de similaridade

n	x	y	Si	n	x	y	Si	n	x	y	Si
1	5	2	0,025	8	0	0,2	0,155	15	0,4	0,4	0,154
2	1,8	0,7	0,073	9	0,6	0,6	0,154	16	0	0,4	0,146
3	0,5	0,2	0,107	10	0	0,6	0,157	17	0,6	0,4	0,139
4	0,2	0,6	0,130	11	0,2	0	0,159	18	0,1	0,1	0,130
5	0	0	0,141	12	0,4	0	0,159	19	0,2	0,4	0,111
6	0,4	0,6	0,150	13	0,3	0,3	0,158	20	0,6	0,2	0,079
7	0,2	0,2	0,155	14	0,6	0	0,156				

### 3. Aplicativos Desenvolvidos

Conforme exposto anteriormente, para evidenciar os dados anômalos porventura existentes em conjunto de dados aplicar-se-á a técnicas Agrupamentos Nebulosos (*Fuzzy clustering*), Máquinas de Vetor Suporte (*Support Vector Machine – one class*) e Função de Similaridade Média.

#### 3.1 Aplicativo de Agrupamento Nebuloso (KPCM)

Apesar do algoritmo ser bastante documentado, não se encontrou aplicativo pronto para suportar os objetivos pretendidos, o que, apesar de demandar esforço de desenvolvimento, permitiu a criação de ferramenta para expor *outliers* de um conjunto de dados, ainda que passando pelo conceito de agrupamento.

O aplicativo apresenta interface amigável, ainda que tenha sido projetada como elemento auxiliar deste estudo, como pode ser observado na figura 15. Requer como entrada dois arquivos texto (dados e protótipo dos *clusters*), com cabeçalho na primeira linha e dados separados por ponto e vírgula. Podem-se alterar os parâmetros do algoritmo, bem como obter um arquivo com os valores iniciais padronizados ou escalonados. Um gráfico de duas dimensões pode ser obtido ao final da execução do algoritmo.

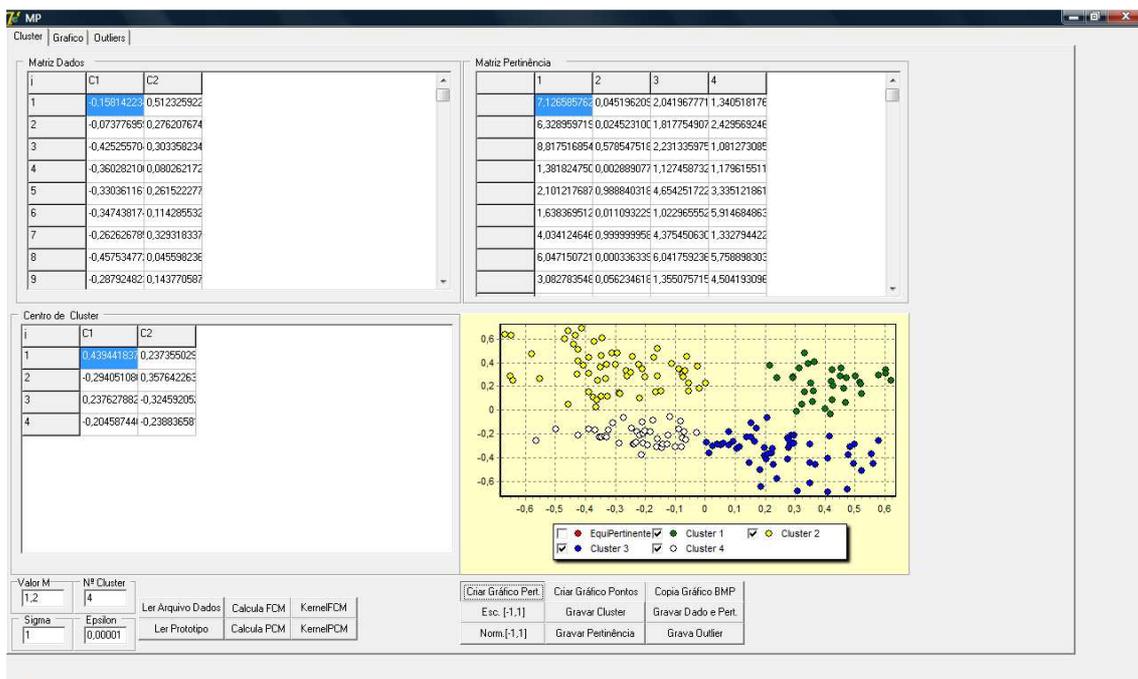


Figura 15. Interface do Aplicativo de Agrupamento Nebuloso



que já se dispunha de aplicativo para escalonar os dados e para construção de gráficos.

Estes programas exigem uma formatação própria para representação dos dados, a começar pelo separador decimal que deve ser o ponto. Como é normal, exige-se como entrada um arquivo texto contendo os dados a serem submetidos. Usualmente, os dados são dispostos em colunas separadas por uma caractere especial (ponto e vírgula, vírgula ou espaço). Nestes programas as dimensões do dado são codificadas com um rótulo, um índice da dimensão e o valor respectivo (<rótulo> <índice 1>:<valor1> <índice 2>:<valor2> ...), como pode ver visto no exemplo com dois atributos exposto na tabela 11:

**Tabela 11.** Formato dos dados LIBSVM

1	1:-0.386046681413173	2:0.0226096422208118
2	1:-0.32941655105485	2:0.00994784208732539
3	1:-0.337286174170709	2:0.0322235776778848
4	1:-0.329884992339841	2:0.00457096878646669
.....		

O aplicativo LIBSVM permite o uso de diversos algoritmos de máquinas de vetor suporte, entre eles, Máquina de Vetor Suporte – Classe Única (*SVM One-class*) (neste caso, o rótulo não tem utilidade), a combinação de diversos tipos de funções kernel, entre eles, o gaussiano, e alteração do parâmetro  $\nu$ , controla a expansão ou retração da margem de separação dos dados.

Na opção *SVM One-class* os programas SVM-TRAIN e SVM-PREDICT seriam executados seqüencialmente e utilizam o mesmo arquivo de dados sem quaisquer alterações, sendo que o segundo utiliza o modelo resultante do primeiro, além dos dados.

Para facilitar seu uso, uma vez que a versão utilizada do aplicativo é para ambiente DOS, construiu-se uma interface de ambiente Windows, que possibilita a formatação dos dados (figura 17), a execução dos programas automaticamente (figura 18) e exibe o gráfico resultante segundo as dimensões escolhidas (figura 19).

N_AIH	VAL_SH	VAL_SP	VAL_SADT	VAL_RN	VAL_ORTP	VAL_SANGLE	VAL_TRANSF	VAL_TOT	VAL_LTI	LIS_TOT
2135065120	-0,98979510	-0,98576625	-0,99779237	-1	-1	-1	-1	-0,99072829	-1	-0,99072957
2135812734	-0,94709755	-0,89592741	-0,99725995	0	-1	-1	-1	-0,94576485	-1	-0,94576558
2136709113	-0,73450827	-0,75669074	-0,92908344	-1	-1	-0,96416596	-1	-0,77496548	-1	-0,77496609
2136711071	-0,94709755	-0,89592741	-0,99725995	0	-1	-1	-1	-0,94576485	-1	-0,94576558
2136713062	-0,94709755	-0,89592741	-0,99725995	0	-1	-1	-1	-0,94576485	-1	-0,94576558
2137592171	-0,96979748	-0,93880610	-0,97697582	-1	-1	-1	-1	-0,96752699	-1	-0,96752675
2137596054	-0,79361424	-0,91289769	-0,90691634	-1	-1	-0,49832349	-1	-0,79878530	-1	-0,79878409
2137596770	-0,94709755	-0,89592741	-0,99725995	0	-1	-1	-1	-0,94576485	-1	-0,94576558
2137596890	-0,96103521	-0,90712574	-0,98353375	-1	-1	-1	-1	-0,95738905	-1	-0,95738972
2137864223	-0,96420065	-0,96366845	-0,97109316	-1	-1	-1	-1	-0,96721089	-1	-0,96721114
2137864718	-0,95593855	-0,92118203	-0,99753265	-1	-1	-1	-1	-0,95796958	-1	-0,95797013
2137864730	-0,93068219	-0,92817749	-0,95414643	-1	-1	-1	-1	-0,93745512	-1	-0,93745536
2137864740	-0,94709755	-0,89592741	-0,99725995	0	-1	-1	-1	-0,94576485	-1	-0,94576558
2137864784	-0,94709755	-0,89592741	-0,99725995	0	-1	-1	-1	-0,94576485	-1	-0,94576558
2137864806	-0,94709755	-0,89592741	-0,99725995	0	-1	-1	-1	-0,94576485	-1	-0,94576558
2137864840	-0,93088364	-0,89592741	-0,99725995	0	-1	-1	-1	-0,93512389	-1	-0,93512304
2137864861	-0,94709755	-0,89592741	-0,99725995	0	-1	-1	-1	-0,94576485	-1	-0,94576558
2137864927	-0,94709755	-0,89592741	-0,99725995	0	-1	-1	-1	-0,94576485	-1	-0,94576558
2137864950	-0,94709755	-0,89592741	-0,99725995	0	-1	-1	-1	-0,94576485	-1	-0,94576558
2137864982	-0,94709755	-0,89592741	-0,99725995	0	-1	-1	-1	-0,94576485	-1	-0,94576558
2137865037	-0,97679294	-0,97803295	-0,98405319	-1	-1	-1	-1	-0,97932900	-1	-0,97933010
2137865070	-0,94709755	-0,89592741	-0,99725995	0	-1	-1	-1	-0,94576485	-1	-0,94576558
2137865136	-0,93088364	-0,89592741	-0,99725995	0	-1	-1	-1	-0,93512389	-1	-0,93512304
2137865235	-0,94709755	-0,89592741	-0,99725995	0	-1	-1	-1	-0,94576485	-1	-0,94576558

Qt Linhas: 29999

Figura 17. Conversão de formato de dados para LIBSVM

**Diretorio dos Aplicativos**  
 C:\UFRJ\apiSVM

**Diretorio do Arquivo de Dados**  
 C:\UFRJ\arqSVM\aih3ESVM.txt

**Diretorio do Resultado**  
 C:\UFRJ\arqRSVM

**Nu**  **C**

**Classe 1**  **Classe -1**

Figura 18. Execução dos programas LIBSVM

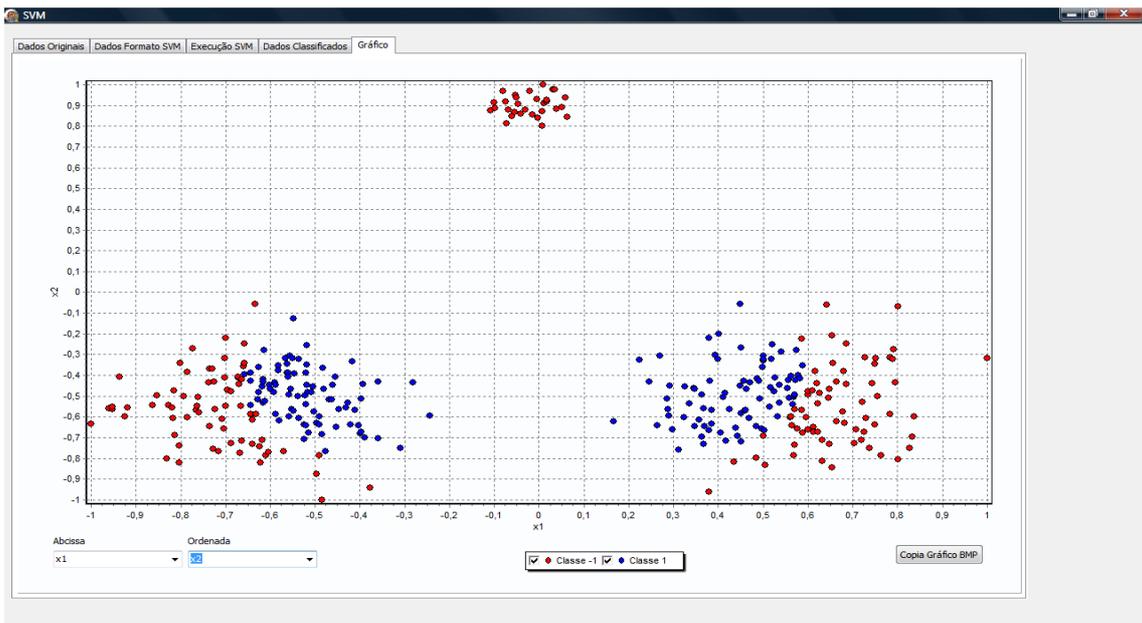


Figura 19. Gráfico do aplicativo LIBSVM

### 3.3 Aplicativo Função de Similaridade Média (FSM)

Projetou-se um aplicativo, que recebendo um arquivo texto com os dados normalizados, calculasse a similaridade média de cada registro conforme exposto anteriormente e fornecesse um resultado ordenado crescentemente desse índice (figura 20). De posse destes, pode-se eleger um ponto de corte que evidencie os registros anômalos (figura 21) e expor resultado em um gráfico segundo as dimensões escolhidas (figura 20).

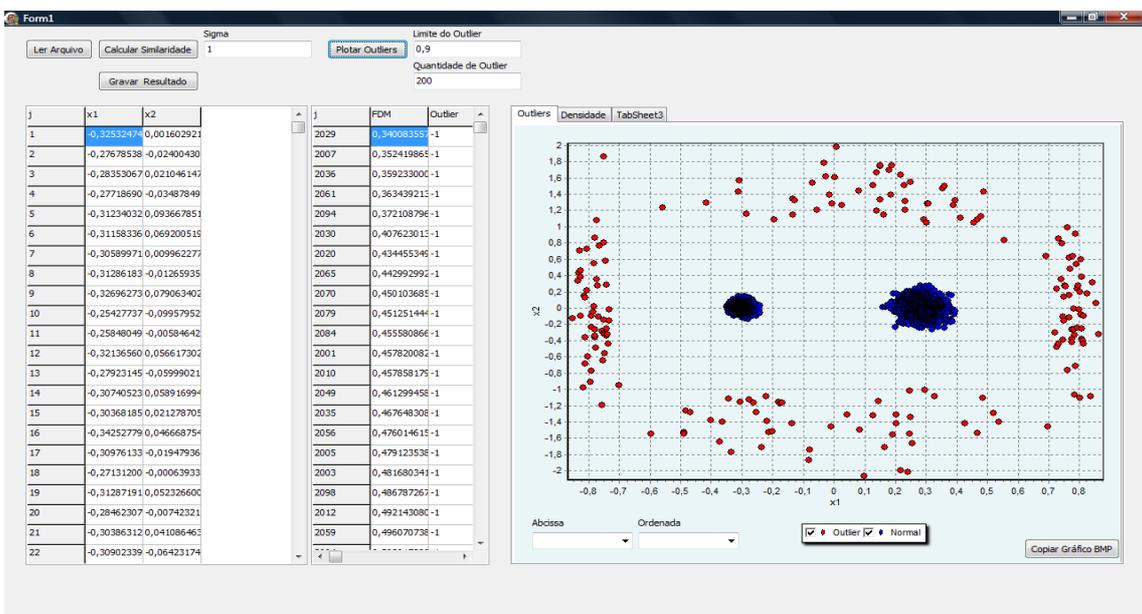


Figura 20. Função de Similaridade Média – apresentação

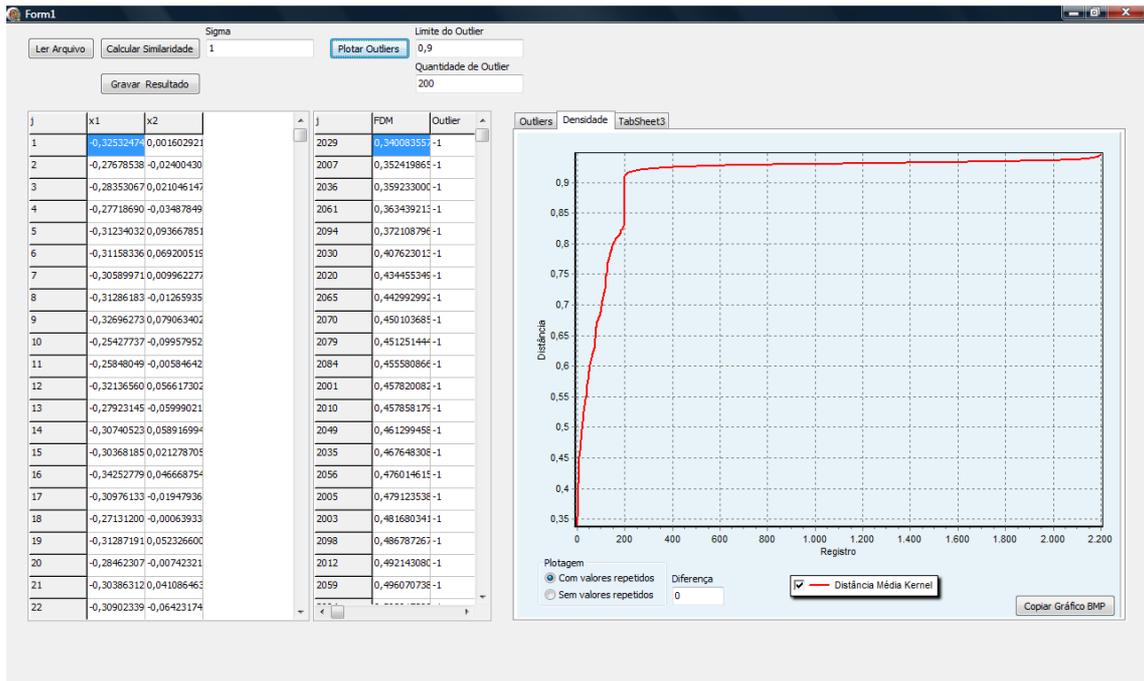


Figura 21. Função de Similaridade Média - Ponto de corte

#### 4. Avaliação dos algoritmos

Antes da aplicação aos dados, há que se aquilatar a correção das ferramentas, que será feita mediante a submissão de dados gerados artificialmente e de dados obtidos em bases de testes.

##### 4.1 – Avaliação do aplicativo de Agrupamento Nebuloso

A fim de avaliar a ferramenta com algoritmo Kernel PCM (KPCM), buscou-se testar sua correção frente a um conjunto de 300 dados gerados artificialmente distribuídos entre dois grupos bem delineados (figura 22), obtendo-se agrupamentos corretos, cujo resultado é exposto na figura 23 abaixo:

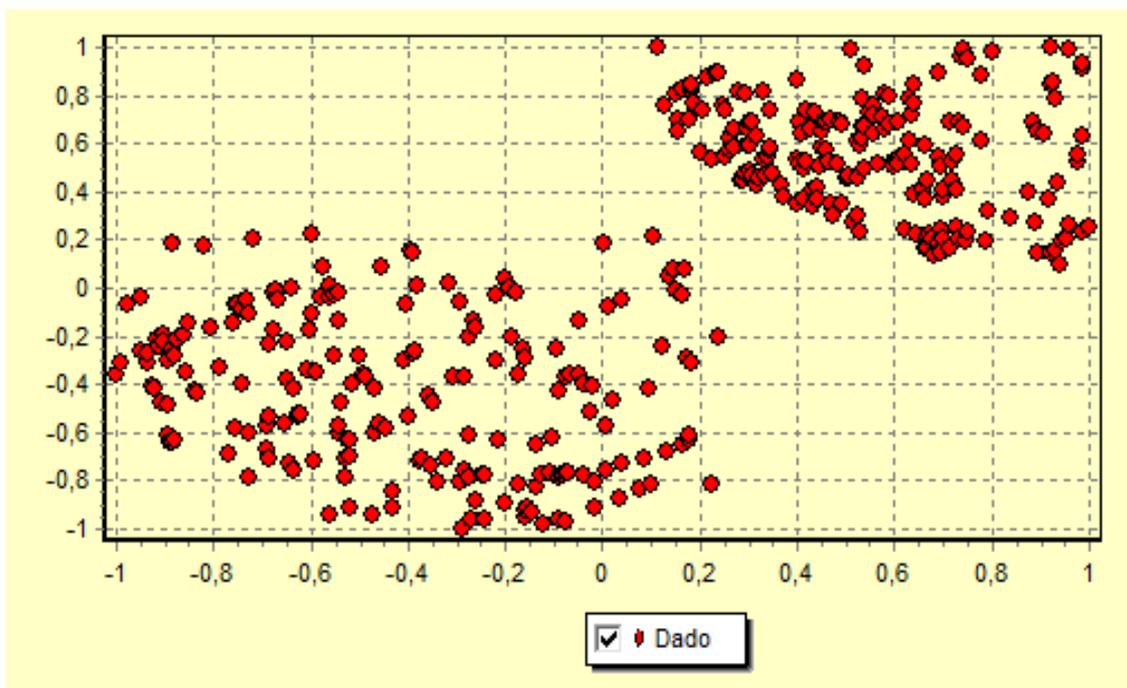


Figura 22 Dados gerados segundo uma distribuição Normal

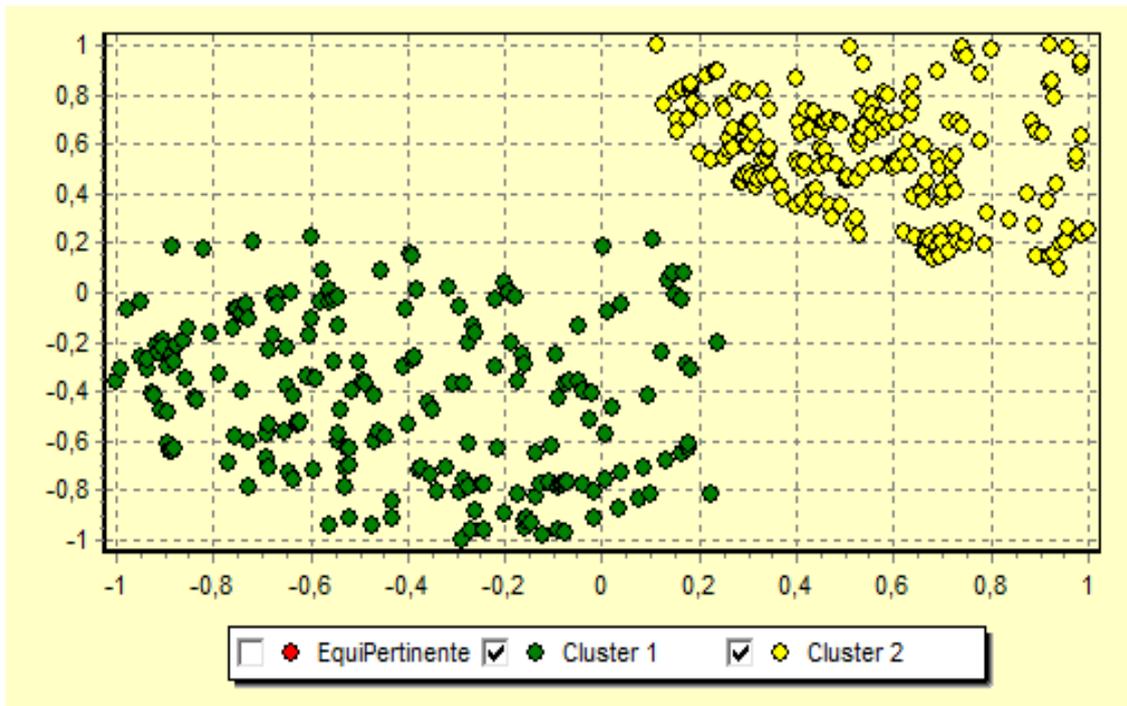


Figura 23 Dados agrupados com algoritmo KPCM

Buscando testar o comportamento do algoritmo frente ao volume de dados, gerou-se 140.000 dados artificiais formando quatro elipses, que foram agrupados corretamente apenas pelo algoritmo usando função kernel, conforme observa-se nas figuras 24 e 25 abaixo:

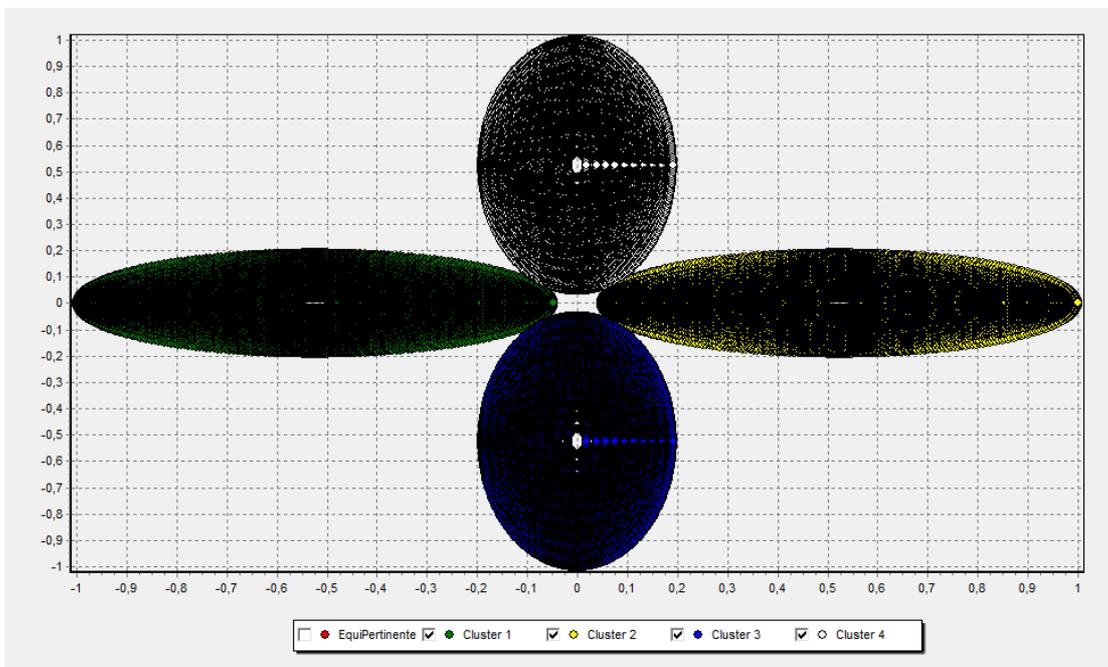


Figura 24 Elipses agrupadas com algoritmo KPCM

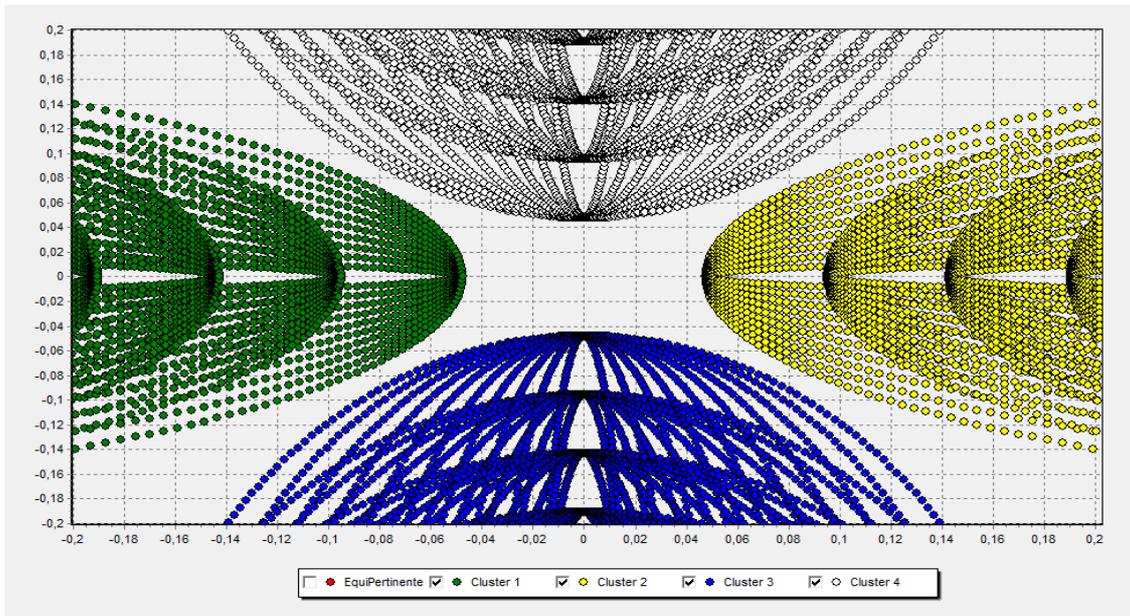


Figura 25 Elipses agrupadas com algoritmo KPCM - ampliação

Testando-se frente a mais de duas dimensões, submeteram-se dados gerados artificialmente de três círculos sobrepostos, e base de dados das flores Iris, e cujos resultados são expostos nas figuras 26, 27 e 28 abaixo:

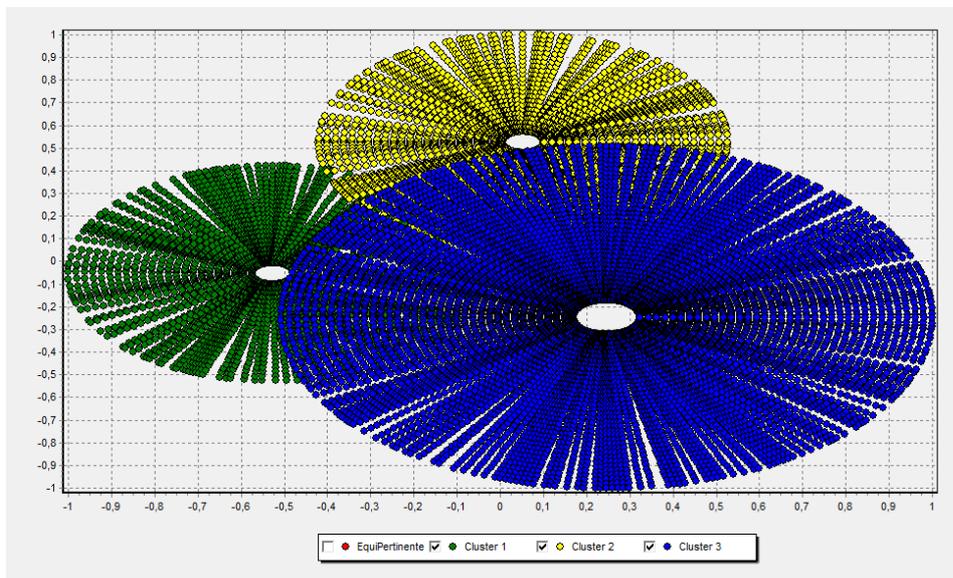


Figura 26 Agrupamento de três círculos (plano XY) - Algoritmo KPCM

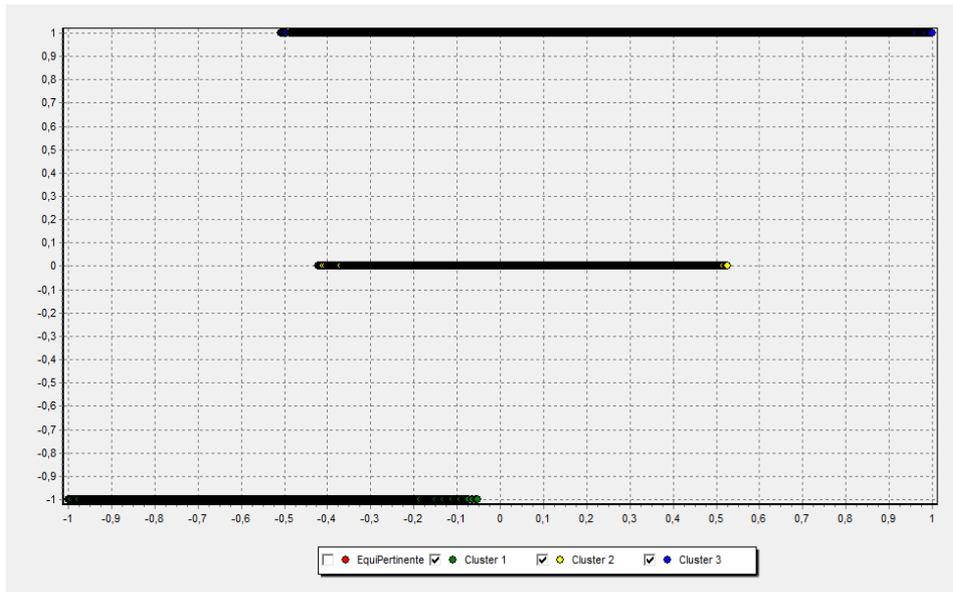


Figura 27 Agrupamento de três círculos (plano XZ) - Algoritmo KPCM

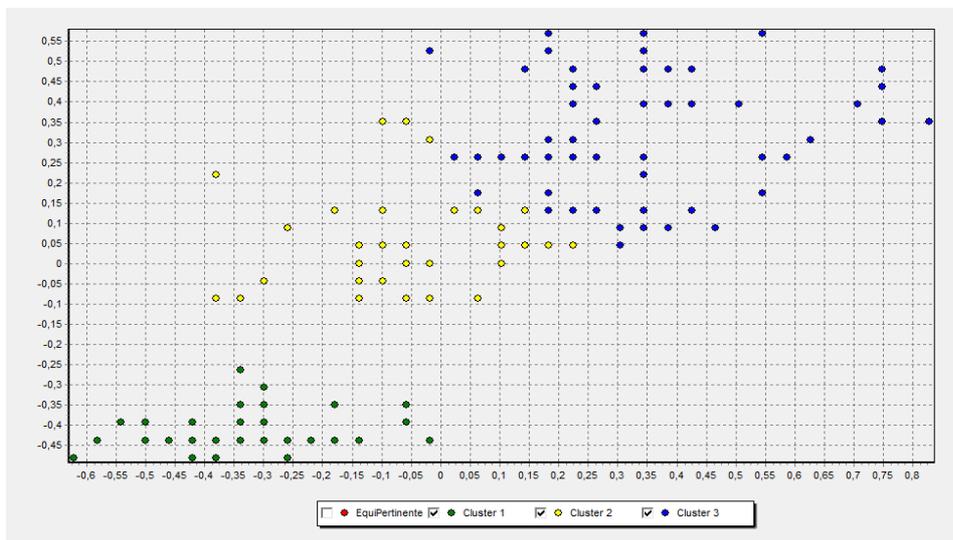
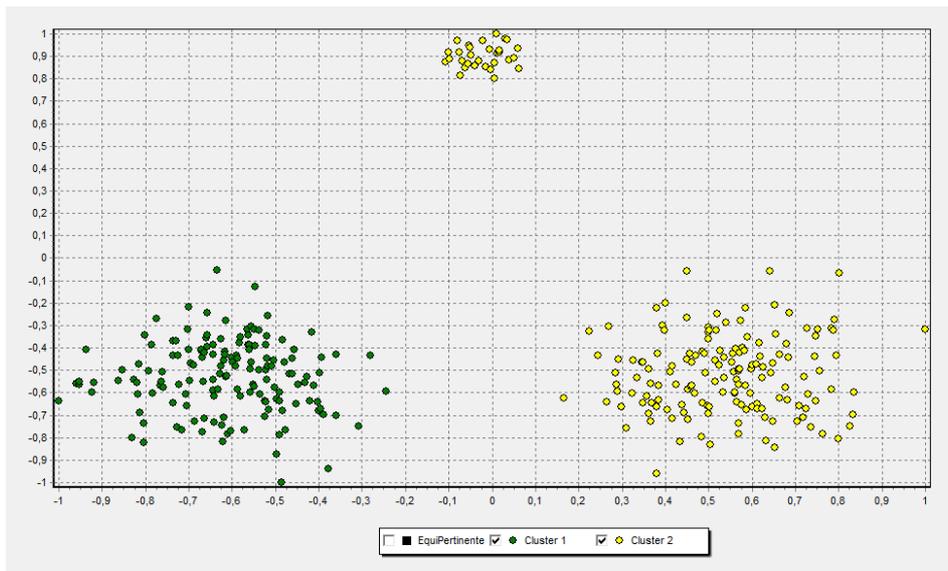
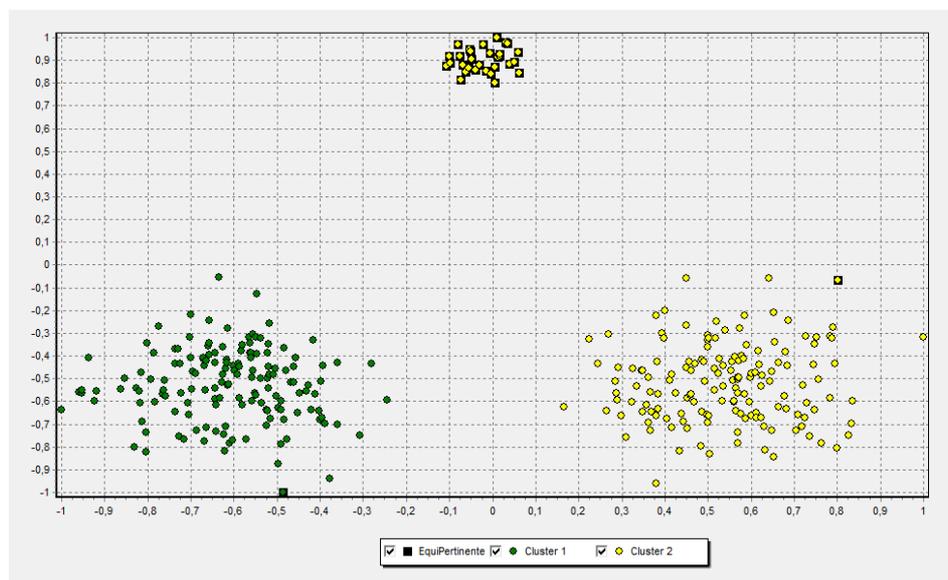


Figura 28 Agrupamento Planta Iris - Algoritmo KPCM

Atestada a correção do algoritmo, passa-se a analisar o foco do estudo, a evidenciação de *outliers* em um conjunto. Para tanto gerou-se uma massa de dados aleatórios, contendo 330 elementos (30 dados anômalos) dispostos conforme figura 29 abaixo :



**Figura 29** Agrupamento de 330 dados gerados artificialmente - Algoritmo KPCM



**Figura 30** Agrupamento de 330 dados gerados artificialmente - Outliers - Algoritmo KPCM

Como exposto anteriormente, o algoritmo PCM tem a capacidade de determinar *outliers* por meio do assinalamento desses dados com um valor de pertinência próximo de zero (OLIVEIRA, J.V.; PEDRYCZ, W., 2007), conforme pode ser visto na figura 30 acima e nos valores das soma das pertinências expostas na tabela 12, onde os registros listados apresentam uma pertinência total baixa, o que permite caracterizá-los como *outliers*.

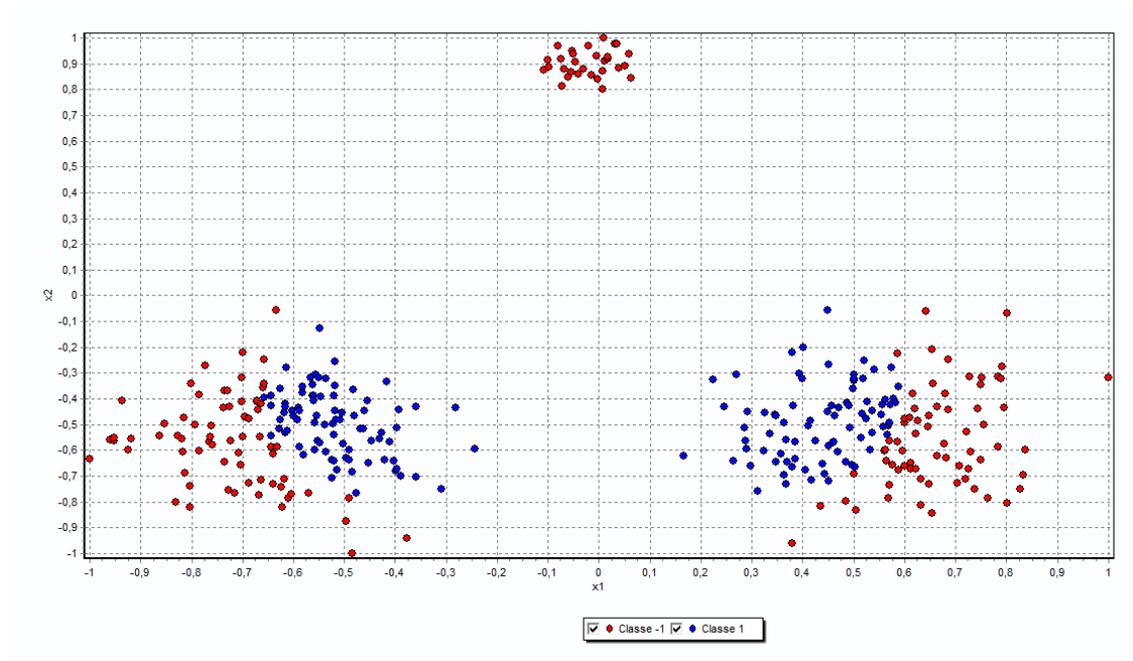
**Tabela 12.** Tabela da soma das pertinências

J	x1	x2	Soma Pertin.	j	x1	x2	Soma Pertin.
317	0,00972	1	1,89E-07	322	-0,10788	0,876741	2,20E-07
310	-0,08071	0,970472	1,94E-07	301	-0,06821	0,878738	2,21E-07
329	0,031391	0,976847	1,95E-07	302	0,050885	0,889993	2,22E-07
327	0,035097	0,975835	1,95E-07	313	-0,02986	0,879809	2,22E-07
304	-0,02059	0,968342	1,96E-07	305	0,039189	0,883195	2,23E-07
318	-0,05291	0,949589	2,00E-07	321	-0,05501	0,867355	2,25E-07
319	-0,05109	0,938683	2,03E-07	330	0,006745	0,871876	2,26E-07
325	-0,00558	0,930011	2,07E-07	309	-0,03989	0,859518	2,29E-07
308	0,058877	0,935729	2,07E-07	316	-0,06163	0,849294	2,32E-07
328	-0,10084	0,915997	2,08E-07	307	-0,0144	0,854282	2,32E-07
324	-0,07504	0,917337	2,08E-07	311	-0,00295	0,839743	2,39E-07
323	0,017471	0,9263	2,09E-07	315	0,062846	0,843662	2,40E-07
326	0,015974	0,917406	2,11E-07	303	-0,07348	0,814392	2,46E-07
306	0,011702	0,911529	2,13E-07	312	0,00644	0,799418	2,58E-07
314	-0,04748	0,904613	2,13E-07	292	-0,48506	-1	9,30E-05
320	-0,09773	0,885622	2,17E-07	257	0,801735	-0,06807	9,69E-05

## 4.2 Avaliação do aplicativo SVM-One Class

Esse algoritmo tem diversas utilidades, mas todas se resumem na tarefa de separar os dados normais dos anômalos. Conforme exposto anteriormente, o parâmetro  $\nu$  controla a taxa de permutação entre o volume da esfera contendo os dados, é a razão entre o volume da esfera e o número de pontos rejeitados. Valores altos para este parâmetro praticamente não influenciam as fronteiras entre as classes, ocorrendo o oposto para valores pequenos (SCHÖLKOPF, B; SMOLA, A. J., 2002) (CHANG, Q. *et al.*, 2007) (CHEN, J. e XU, G., 2009).

Assim, utilizando a mesma massa de dados exposta na figura 30 anterior e a submetendo ao aplicativo, chega-se a separar o conjunto em 164 registros na classe 1 e 166 na classe -1, conforme pode ser visto na figura 31:



**Figura 31** Separação de dados artificiais - Algoritmo SVM *One-class* ( $\nu = 0,5$ )

Verifica-se que o algoritmo consegue evidenciar os dados sabidamente anômalos, captando, entretanto, outros considerados normais.

Esse algoritmo tende a buscar uma esfera que conterà a maior parte dos dados, os ditos normais (classificados como classe 1), restando na parte externa os dados anormais (classificados como classe -1). O equilíbrio entre o raio da hiperesfera e, conseqüentemente, o número de dados anômalos é determinado pelo valor inserido no parâmetro  $\nu$ , cujo incremento ou decremento levará a diminuir ou aumentar o raio (aumentando ou diminuindo a quantidade de outliers) (CHANG, Q. *et al.*,2007).

Conforme exposto por CHANG, Q. *et al.* (2007) e BEN-HUR, A. *et al* (2001), o parâmetro  $\nu$  deve ser superior a  $1/m$ , onde  $m$  é a quantidade de registros em análise, caso contrário, a maior parcela de outliers será confundida com dados normais. Entretanto, vários aplicativos desenvolvidos selecionam automaticamente este valor limite, como no caso do LIBSVM, que assume tal valor como padrão, se outro não for atribuído a esta variável.

Esta propriedade de movimentar a fronteira entre as duas classes fica claramente exposta ao se submeter ao algoritmo o conjunto de 330 dados gerados artificialmente com o subconjunto de *outliers* visto claramente. Assim, submetendo-os ao algoritmo com valores decrescentes para o parâmetro  $\nu = 0,3$  (figura 32) ;  $0,03$  (figura 33), e o limite  $\nu = 1/m = 1/330 = 0,003$ , onde se comprova a afirmação acima, posto que quase não há registros classificados como classe -1, conforme vê-se na figura 34 abaixo:

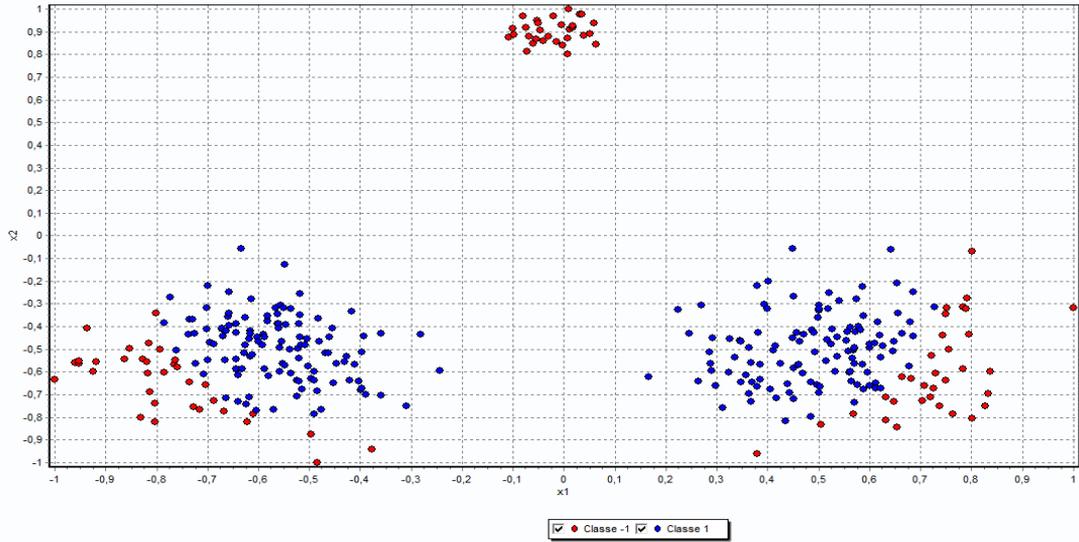


Figura 32 Separação de dados artificiais - Algoritmo SVM One-class ( $u = 0,3$ )

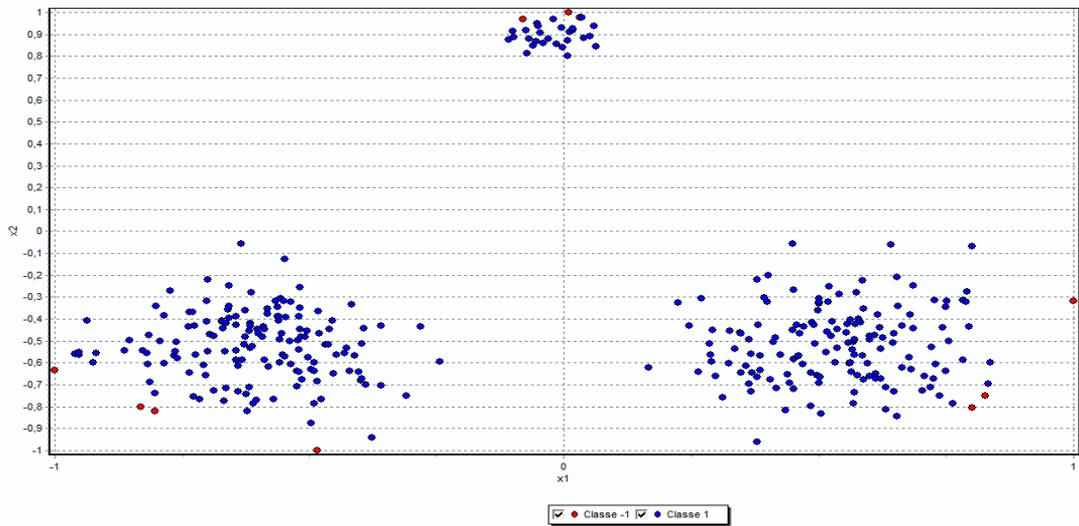


Figura 33 Separação de dados artificiais - Algoritmo SVM One-class ( $u = 0,03$ )

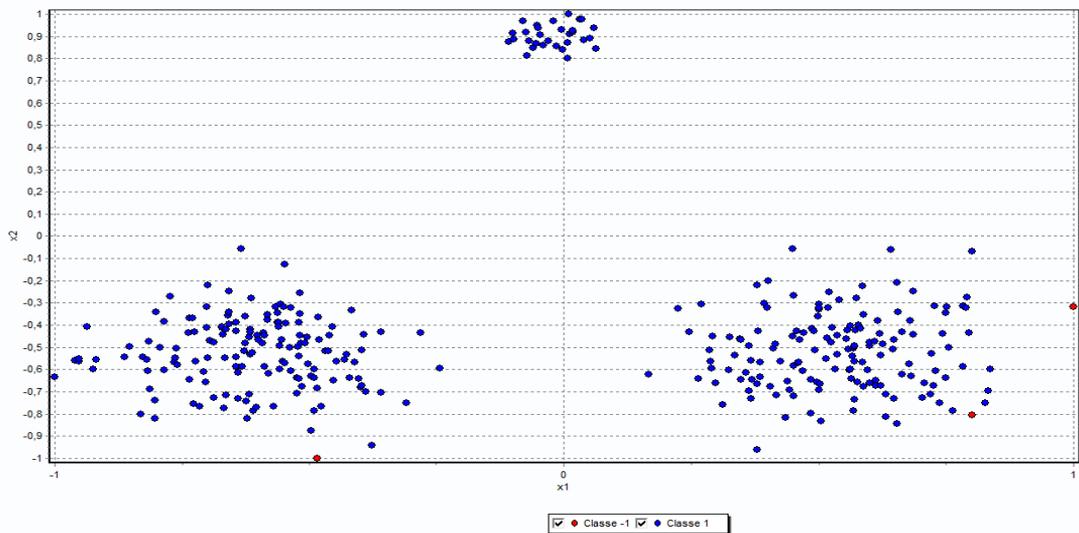
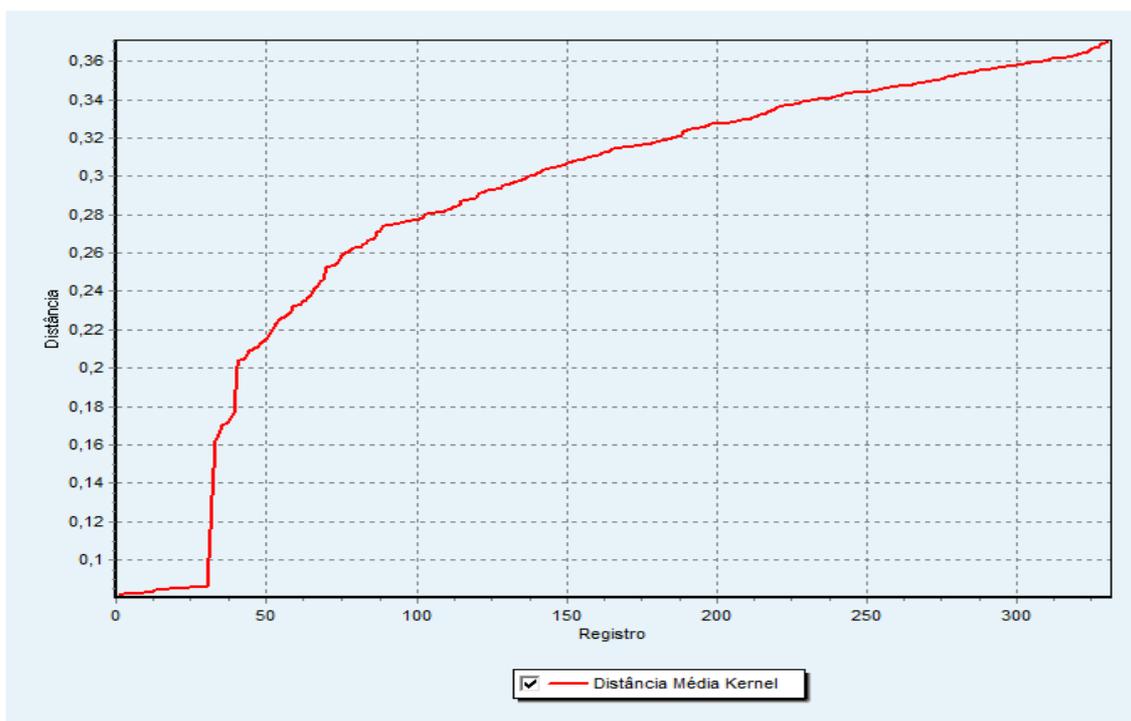


Figura 34 Separação de dados artificiais - Algoritmo SVM One-class ( $u = 0,003$ )

Assim, apesar de se saber que o parâmetro  $\nu \in (0,1]$ , pode-se restringir seu intervalo para  $\nu \in (1/m, 1]$ , para um conjunto de  $m$  dados. O intervalo menor admite ainda escolhas diversas, o que mantém a afirmativa de TRAN, Q. *et al.* (2003) que a escolha de  $\nu$  depende da arte e experiência do pesquisador.

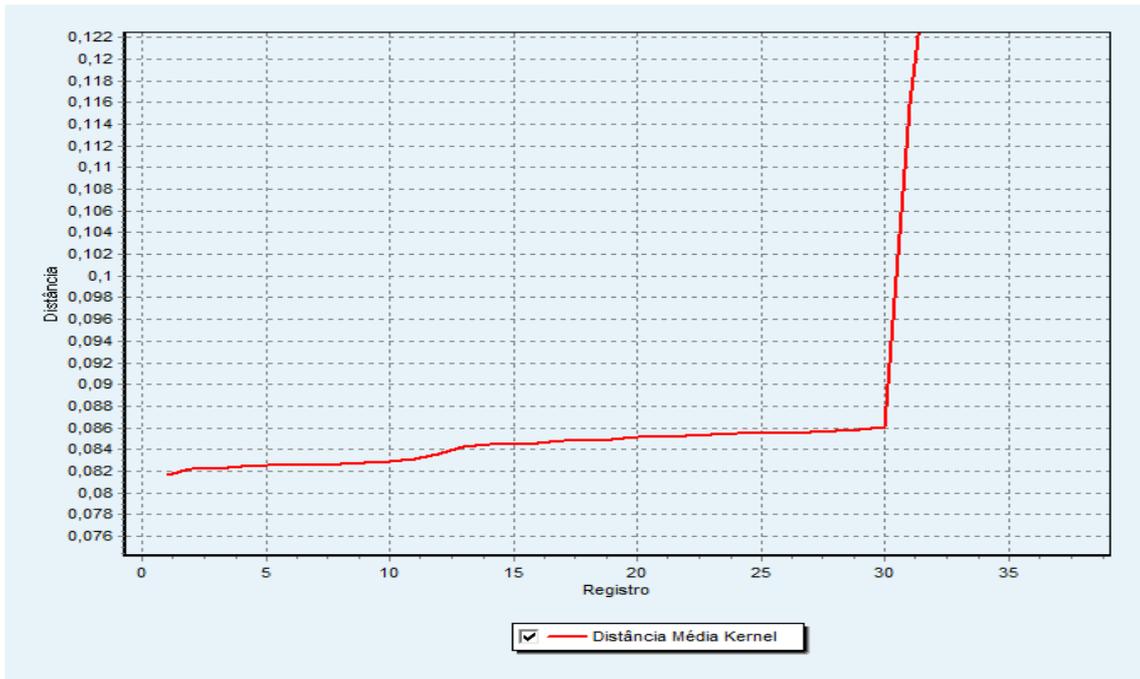
### 4.3 Avaliação do aplicativo de Função de Distância Média

Utilizando a mesma massa de teste, calcula-se a similaridade média de cada ponto a todos os demais. Aplicando-se neste cálculo uma função kernel gaussiana, obtém-se uma série de distâncias totais médias, que ordenados ascendentemente, apresentam-se conforme a figura 35 abaixo:



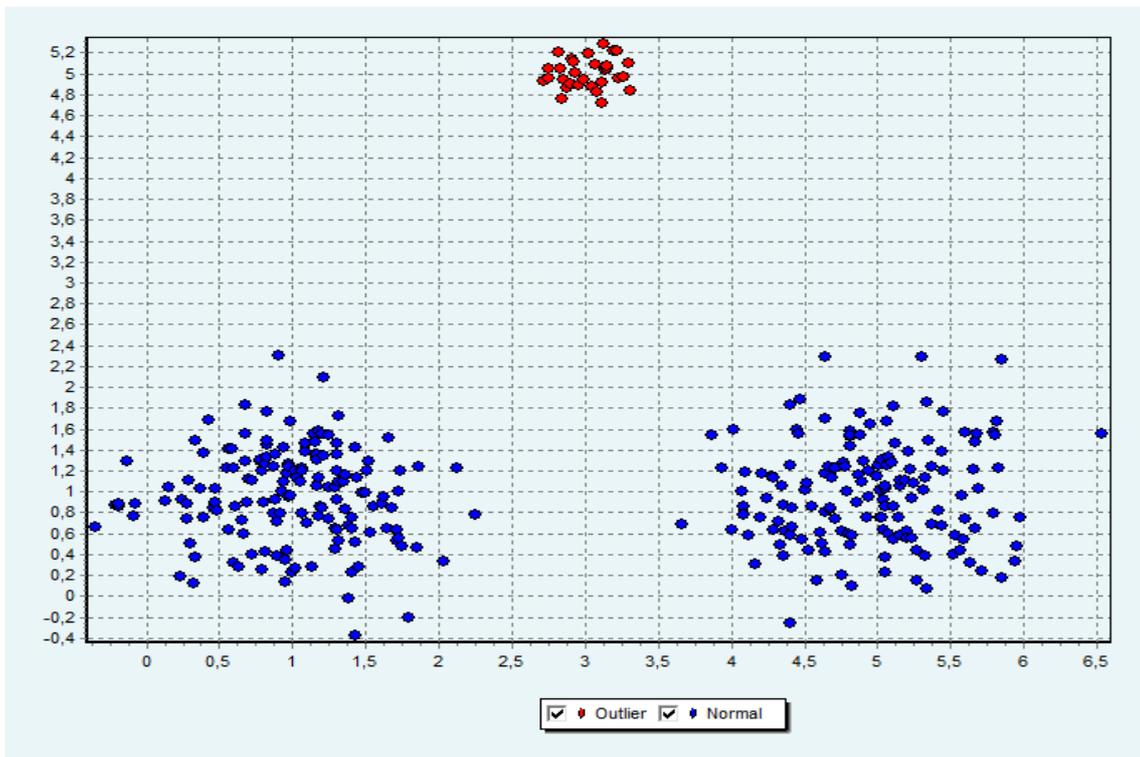
**Figura 35** Função de Similaridade Média – ponto de corte

Ora, como os valores anômalos têm uma medida de similaridade mais baixa, há que se verificar quais pontos são esses. Ampliando-se, observa-se o exposto na figura 36 abaixo, onde se observa que os primeiros trinta elementos apresentam um índice baixo,  $S = 0,086$ , e há uma inflexão na curva, caracterizando uma mudança de comportamento.



**Figura 36** Função de Similaridade Média – ponto de corte (Ampliação)

Utilizando este valor de similaridade como fronteira entre os dados normais e *outliers*, tem-se a seguinte separação entre os dados vista na figura 37:



**Figura 37** Função de distância Média - outliers

Desses testes, concluí-se que os algoritmos são efetivos para elencar os dados divergentes. Os dados colocados como intencionalmente *outliers* foram

evidenciados por todos os algoritmos, com pequenas variações decorrentes de seus parâmetros de entrada.

#### 4.4 Precisão dos Algoritmos

Determinar a acurácia e precisão do algoritmo em métodos não-supervisionados é algo inviável, pois não há como verificar o grau de acerto quando não existem registros para comparação. Entretanto, podem-se utilizar elementos de aprendizado supervisionados para testar a validade do algoritmo, valendo-se do truque proposto por CHANG, Q. *et al.* (2007), onde se utiliza uma base de dados criada artificialmente, contendo dados normais e anômalos, sendo que o número do primeiro será significativamente maior que o dos *outliers*.

Para tal, foram utilizadas três bases de dados disponibilizadas por *UCI Machine Learning Repository* e obtidas em <http://www.ics.uci.edu/~mllearn/MLRepository.html>, acrescentando-se alguns registros anômalos com valores aleatórios superiores ao máximo do conjunto (apenas uma dimensão foi alterada):

- Planta Íris, com um conjunto de 150 registros e quatro dimensões, onde se acrescentou 10 registros anômalos para esta aferição.
- Reconhecimento de vinhos, uma base de dados com 178 registros e 13 dimensões, onde acrescentou-se 20 *outliers*.
- Dados do molusco Abalone, constituído de 4.177 registros e 8 atributos, onde se acrescentou 28 *outliers*.

Há duas classes para separar os dados e pode ocorrer classificação errada ou correta, o que acarreta quatro possibilidades de resultados:

- a) o registro pertencente à classe 1 e é corretamente classificado. Este resultado é denominado positivo verdadeiro (TP)
- b) o registro pertencente à classe 1 e é classificado como da outra classe. Denomina-se tal ocorrência como falso negativo (FN)
- c) o registro pertence à classe -1 e é classificado corretamente. O resultado é intitulado negativo verdadeiro (TN)
- d) o registro é da classe -1 e é caracterizado como pertencente à classe 1. Tal é definido como resultado falso positivo (FP).

Estas definições dão origem à matriz de confusão, onde se relaciona as quantidades de registros classificados correta e incorretamente, conforme vê-se na tabela 13 abaixo (FAWCETT , T., 2005) (LASKO, T. *et al.*, 2005) :

Tabela 13. Matriz de confusão

		Classe Verdadeira	
		Classe 1	Classe -1
Classificação Ocorrida	Classe 1	positivo verdadeiro (TP)	falso positivo (FP)
	Classe -1	falso negativo (FN)	negativo verdadeiro (TN)

Dessa matriz extrai-se quatro métricas para avaliar o resultado da classificação:

- a) acurácia, que indica a razão dos dados classificados corretamente, assim definido:

$$a = \frac{\text{número de registros TP} + \text{número de registros TN}}{\text{número Total de registros}}$$

- b) precisão, que indica a capacidade de detecção da classe 1:

$$p = \frac{\text{número de registros TP}}{\text{número de registros TP} + \text{número de registros FP}}$$

- c) taxa de acerto, proporção da classificação correta na classe 1:

$$ta = \frac{\text{número de registros TP}}{\text{número de registros Classe 1}}$$

- d) taxa de falso alarme, proporção classificação incorreta na classe -1:

$$fa = \frac{\text{número de registros FP}}{\text{número de registros Classe - 1}}$$

#### 4.4.1 Precisão do aplicativo de Agrupamento Nebuloso (KPCM)

Conforme já exposto, este algoritmo determina o grau de pertencimento do dado aos cluster elencados, sendo que os dados anômalos terão o somatório de suas pertinências com valores muito pequenos.

Neste aplicativo, utiliza-se também a função kernel gaussiana com uma equação semelhante a do algoritmo SVM-ONE CLASS implementado no LIBSVM, como exposto abaixo:

- LIBSVM

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \text{ onde } \gamma = 1/\sigma$$

- KPCM

$$K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma)$$

Os parâmetros  $\gamma$  e  $\sigma$  têm a mesma função, mas foram dispostos de forma inversa nos aplicativos, sendo os seguintes valores a serem usados:

LIBSVM	KPCM
$\gamma$	$\sigma$
2	0,25
1	0,5
0,5	1
0,25	2
0,125	4

Submetendo-se a base de dados da Planta Íris a este algoritmo, variando-se o valor do parâmetro c, alcança-se a seguinte matriz de confusão (tabela 14) e índices de precisão exposto na tabela 15:

Tabela 14. Matriz de confusão - KPCM

$\sigma$	0,25		0,5		1	
	Normal	Anômalo	Normal	Anômalo	Normal	Anômalo
Normal	150	0	150	0	150	0
Anômalo	0	10	0	10	0	10
Total	150	10	150	10	150	10

$\sigma$	2		4	
	Normal	Anômalo	Normal	Anômalo
Normal	150	0	150	0
Anômalo	0	10	0	10
Total	150	10	150	10

Tabela 15. Índices de precisão KPCM

Índice	$\sigma$				
	0,25	0,5	1	2	4
Acurácia	1,00	1,00	1,00	1,00	1,00
Precisão	1,00	1,00	1,00	1,00	1,00
Taxa de Acerto	1,00	1,00	1,00	1,00	1,00
Taxa Falso alarme	0,00	0,00	0,00	0,00	0,00

Observa-se que os dados anômalos podem ser evidenciados por meio da pertinência total exposto na figura 38, onde se verifica que há uma mudança de tendência no décimo registro e resultam na indicação dos *outliers* exibidos na figura 39

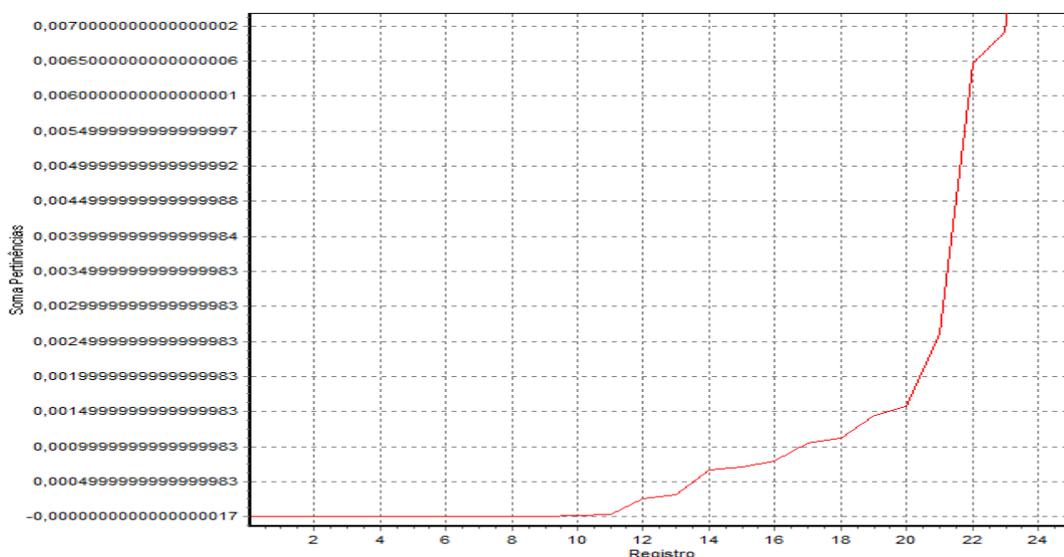


Figura 38 - Pertinência total – Planta Íris - Algoritmo KPCM

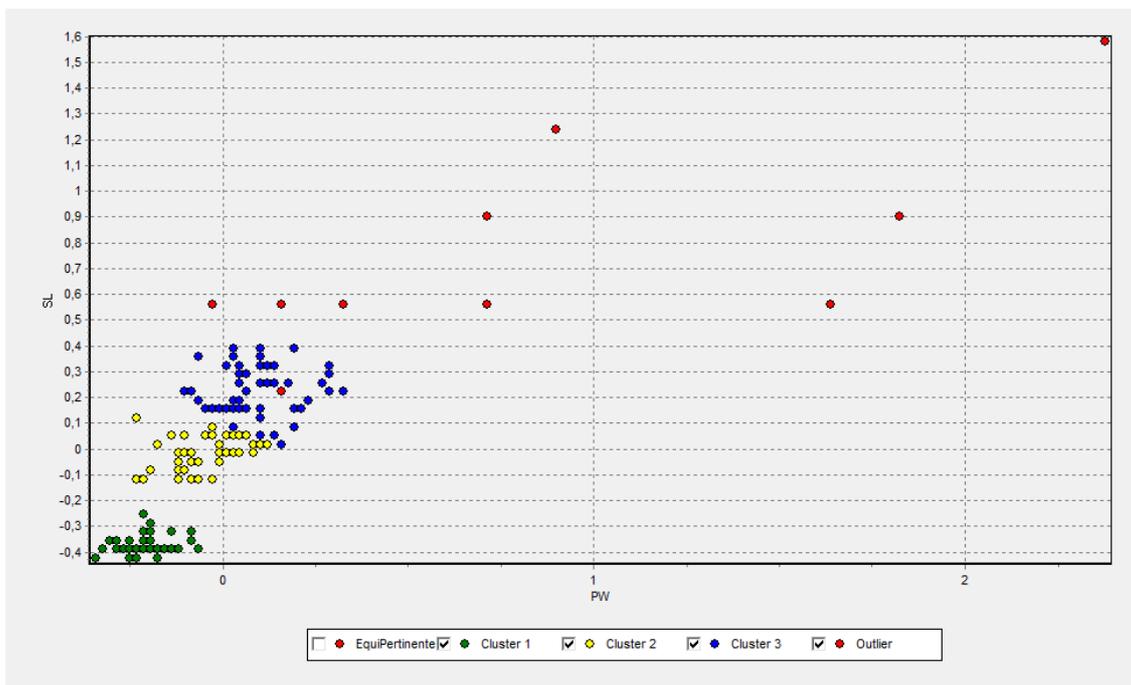


Figura 39 KPCM – Outliers – Planta Íris (PW – Petal Width SL – Sepal Length)

Submetendo-se a base de dados de reconhecimento de vinhos a este algoritmo obtém-se a matriz de confusão exposta na tabela 16 e índices de precisão na tabela 17 :

Tabela 16. Matriz de Confusão KPCM - Reconhecimento de Vinhos

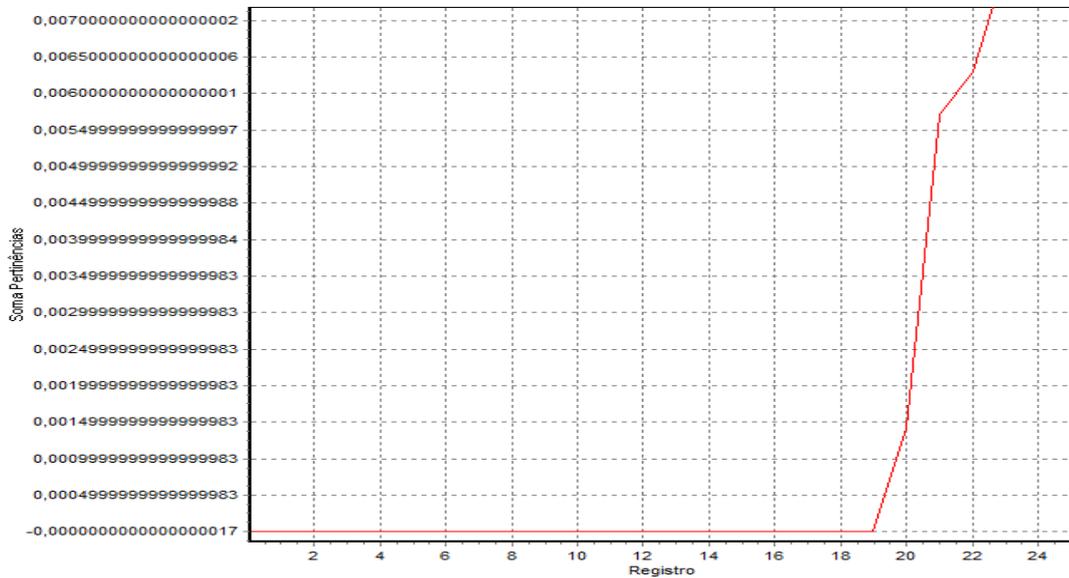
$\sigma$	0,25		0,5		1	
	Normal	Anômalo	Normal	Anômalo	Normal	Anômalo
Normal	178	1	178	1	178	1
Anômalo	0	19	0	19	0	19
Total	178	20	178	20	178	20

$\sigma$	2		4	
	Normal	Anômalo	Normal	Anômalo
Normal	178	1	178	1
Anômalo	0	19	0	19
Total	178	20	178	20

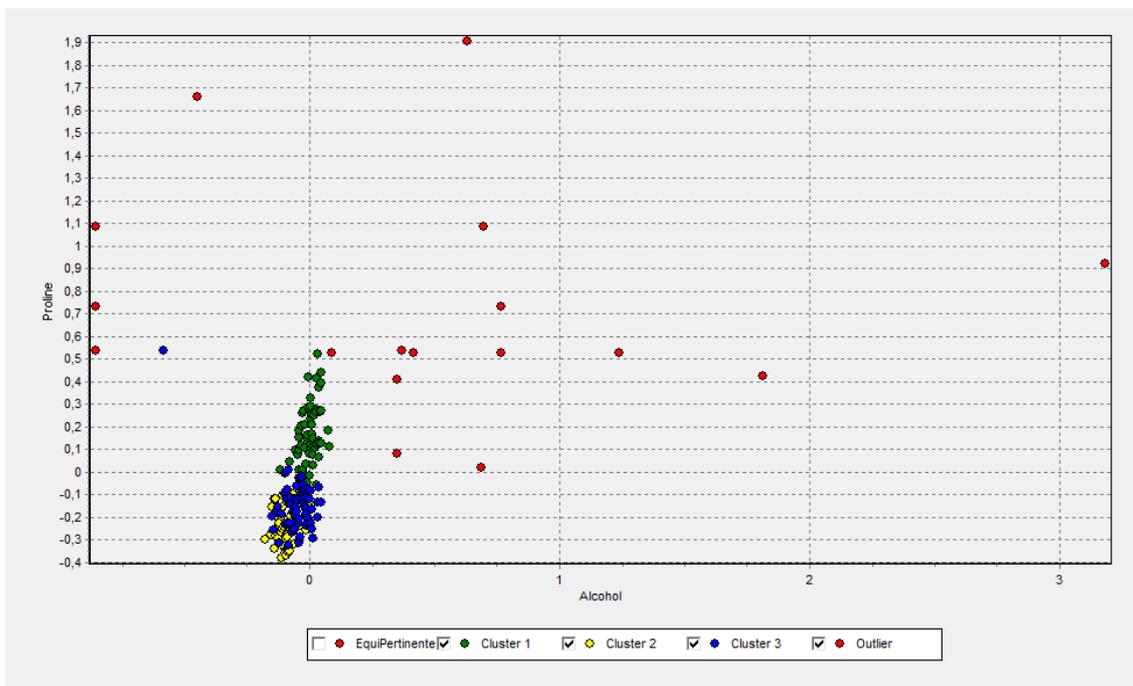
Tabela 17. Índices de precisão - KPCM - Reconhecimento de Vinhos

Índice	$\sigma$				
	0,25	0,5	1	2	4
Acurácia	0,99	0,99	0,99	0,99	0,99
Precisão	0,99	0,99	0,99	0,99	0,99
Taxa de Acerto	1,00	1,00	1,00	1,00	1,00
Taxa Falso alarme	0,05	0,05	0,05	0,05	0,05

De forma semelhante ao anterior, consegue-se separar os *outliers* existentes na base de dados conforme se vê na figura 40, que exibe a curva de pertinências totais, e agrupamento exposto na figura 41 :



**Figura 40** KPCM - Pertinências totais - Reconhecimento de Vinhos



**Figura 41** KPCM - Evidenciação dos outliers - Reconhecimento de Vinhos

O algoritmo KPCM consegue separar o subconjunto dos outliers dos demais dados do conjunto Abalone, estabelecendo a seguinte matriz de confusão (tabela 18) e índices de precisão (tabela 19):

**Tabela 18.** KPCM - Matriz de confusão - Abalones

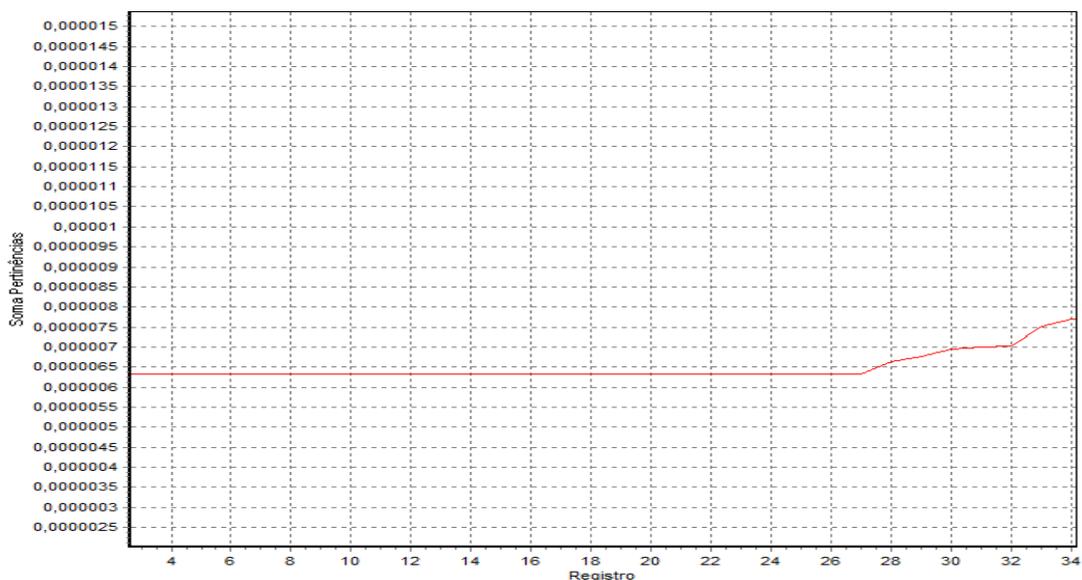
$\sigma$	0,25		0,5		1	
	Normal	Anômalo	Normal	Anômalo	Normal	Anômalo
Normal	4.177	1	4.177	1	4.177	1
Anômalo	0	27	0	27	0	27
Total	4.177	28	4.177	28	4.177	28

$\sigma$	2		4	
	Normal	Anômalo	Normal	Anômalo
Normal	4.177	1	4.177	1
Anômalo	0	27	0	27
Total	4.177	28	4.177	28

**Tabela 19.** KPCM - Índices de precisão - Abalones

Índice	$\sigma$				
	0,25	0,5	1	2	4
Acurácia	1,00	1,00	1,00	1,00	1,00
Precisão	1,00	1,00	1,00	1,00	1,00
Taxa de Acerto	1,00	1,00	1,00	1,00	1,00
Taxa Falso alarme	0,04	0,04	0,04	0,04	0,04

A submissão dos dados do conjunto Abalone apresenta o ponto de corte evidenciado pela curva de pertinências totais, conforme figura 42, e evidencia os outliers como se observa na figura 43:



**Figura 42** KPCM - Pertinência total - Abalones

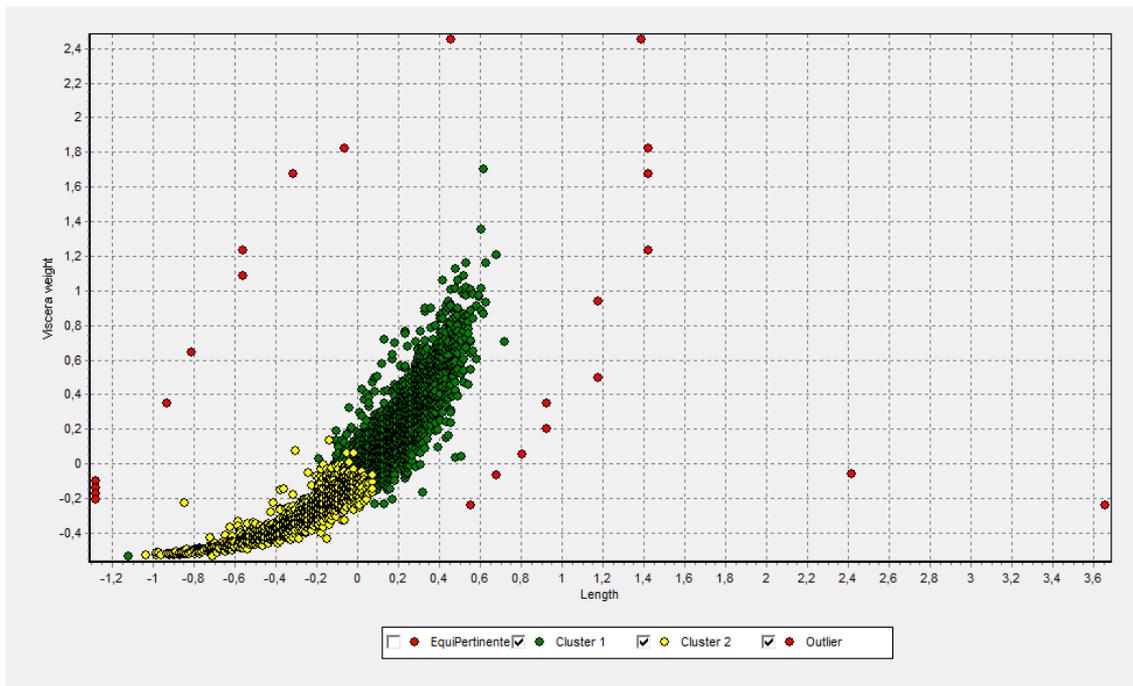


Figura 43 KPCM - Evidenciação de *Outliers* - Abalones

#### 4.4.2 Precisão do aplicativo SVM-One Class do aplicativo LIBSVM

Submetendo-se os valores do conjunto de dados da Planta Íris, variando-se os valores dos parâmetros  $c$  e  $\sigma$ , obtém-se a matriz de confusão exposta na tabela 20 e precisão da tabela 21 :

Tabela 20. SVM - One Class - Matriz de Confusão - Planta Iris

C	$\nu$	0,5		0,25		0,125	
		CLASSE 1	CLASSE -1	CLASSE 1	CLASSE -1	CLASSE 1	CLASSE -1
1	CLASSE 1	81	0	119	0	140	1
	CLASSE -1	69	10	31	10	10	9
	Total	150	10	150	10	150	10
0,5	CLASSE 1	81	0	119	0	138	1
	CLASSE -1	69	10	31	10	12	9
	Total	150	10	150	10	150	10
0,25	CLASSE 1	80	0	122	0	138	1
	CLASSE -1	70	10	28	10	12	9
	Total	150	10	150	10	150	10
0,125	CLASSE 1	81	0	121	0	140	1
	CLASSE -1	69	10	29	10	10	9
	Total	150	10	150	10	150	10
0,0625	CLASSE 1	81	0	121	0	138	1
	CLASSE -1	69	10	29	10	12	9
	Total	150	10	150	10	150	10

C	v	0,0625		0,0313	
		CLASSE 1	CLASSE -1	CLASSE 1	CLASSE -1
1	CLASSE 1	148	3	147	7
	CLASSE -1	2	7	3	3
	Total	150	10	150	10
0,5	CLASSE 1	146	3	146	7
	CLASSE -1	4	7	4	3
	Total	150	10	150	10
0,25	CLASSE 1	145	5	148	6
	CLASSE -1	5	5	2	4
	Total	150	10	150	10
0,125	CLASSE 1	145	5	147	7
	CLASSE -1	5	5	3	3
	Total	150	10	150	10
0,0625	CLASSE 1	146	5	148	7
	CLASSE -1	4	5	2	3
	Total	150	10	150	10

Tabela 21. SVM -One Class - Índices de Precisão - Planta Íris

c	Índice	v				
		0,5	0,25	0,125	0,0625	0,0313
1	Acurácia	0,57	0,81	0,93	0,97	0,94
	Precisão	1,00	1,00	0,99	0,98	0,95
0,5	Acurácia	0,57	0,81	0,92	0,96	0,93
	Precisão	1,00	1,00	0,99	0,98	0,95
0,25	Acurácia	0,56	0,83	0,92	0,94	0,95
	Precisão	1,00	1,00	0,99	0,97	0,96
0,125	Acurácia	0,57	0,82	0,93	0,94	0,94
	Precisão	1,00	1,00	0,99	0,97	0,95
0,0625	Acurácia	0,57	0,82	0,92	0,94	0,94
	Precisão	1,00	1,00	0,99	0,97	0,95
1	Taxa de Acerto	0,54	0,79	0,93	0,99	0,98
	Taxa Falso alarme	0,00	0,00	0,10	0,30	0,70
0,5	Taxa de Acerto	0,54	0,79	0,92	0,97	0,97
	Taxa Falso alarme	0,00	0,00	0,10	0,30	0,70
0,25	Taxa de Acerto	0,53	0,81	0,92	0,97	0,99
	Taxa Falso alarme	0,00	0,00	0,10	0,50	0,60
0,125	Taxa de Acerto	0,54	0,81	0,93	0,97	0,98
	Taxa Falso alarme	0,00	0,00	0,10	0,50	0,70
0,0625	Taxa de Acerto	0,54	0,81	0,92	0,97	0,99
	Taxa Falso alarme	0,00	0,00	0,10	0,50	0,70

O aplicativo evidencia os *outliers* conforme figura 44, com  $\sigma = 1$  e  $\nu = 0,0625$ :

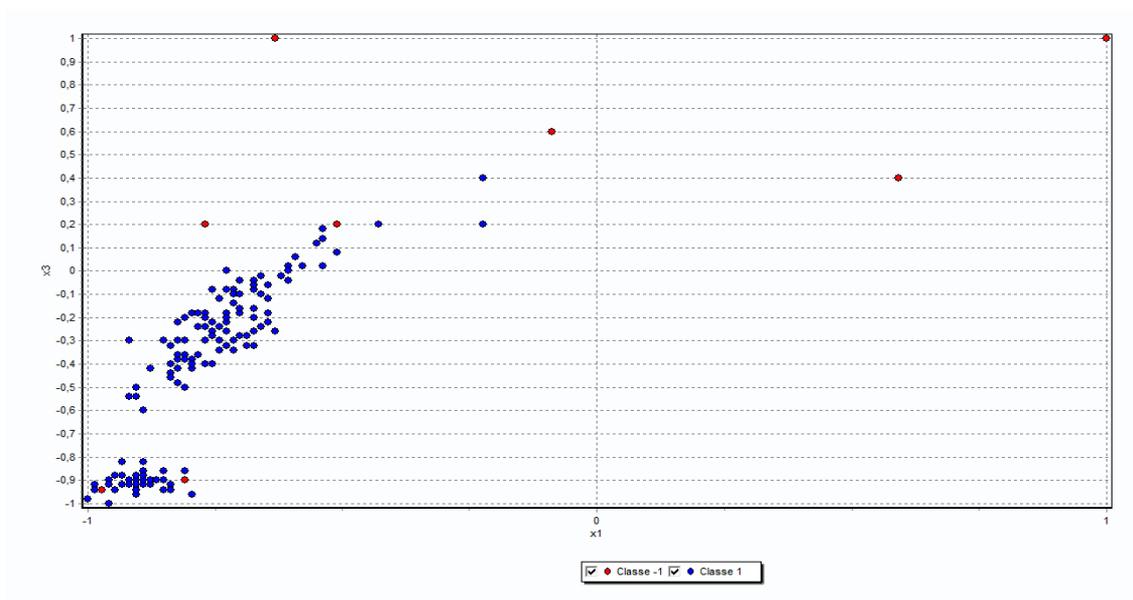


Figura 44 SVM-One Class - Outliers - Planta Íris

O aplicativo apresenta a matriz de confusão exposta na tabela 22 e precisão descrita na tabela 23, quando se submete a base de dados de Reconhecimento de Vinhos:

Tabela 22. SVM - One Class - Matriz de confusão - Reconhecimento de Vinhos

C	$\nu$	0,5		0,25		0,125	
		CLASSE 1	CLASSE -1	CLASSE 1	CLASSE -1	CLASSE 1	CLASSE -1
1	CLASSE 1	100	0	149	0	170	3
	CLASSE -1	78	20	29	20	8	17
	Total	178	20	178	20	178	20
0,5	CLASSE 1	98	0	148	0	170	4
	CLASSE -1	80	20	30	20	8	16
	Total	178	20	178	20	178	20
0,25	CLASSE 1	99	0	149	0	168	6
	CLASSE -1	79	20	29	20	10	14
	Total	178	20	178	20	178	20
0,125	CLASSE 1	98	0	148	0	167	7
	CLASSE -1	80	20	30	20	11	13
	Total	178	20	178	20	178	20
0,0625	CLASSE 1	98	0	147	0	165	9
	CLASSE -1	80	20	31	20	13	11
	Total	178	20	178	20	178	20

C	v	0,0625		0,0313	
		CLASSE 1	CLASSE -1	CLASSE 1	CLASSE -1
1	CLASSE 1	174	10	176	15
	CLASSE -1	4	10	2	5
	Total	178	20	178	20
0,5	CLASSE 1	174	10	176	15
	CLASSE -1	4	10	2	5
	Total	178	20	178	20
0,25	CLASSE 1	174	12	176	15
	CLASSE -1	4	8	2	5
	Total	178	20	178	20
0,125	CLASSE 1	174	12	177	16
	CLASSE -1	4	8	1	4
	Total	178	20	178	20
0,0625	CLASSE 1	172	12	175	16
	CLASSE -1	6	8	3	4
	Total	178	20	178	20

Tabela 23. SVM - One Class - Índices de Precisão - Verificação de Vinho

c	Índice	v				
		0,5	0,25	0,125	0,0625	0,0313
1	Acurácia	0,61	0,85	0,94	0,93	0,91
	Precisão	1,00	1,00	0,98	0,95	0,92
0,5	Acurácia	0,60	0,85	0,94	0,93	0,91
	Precisão	1,00	1,00	0,98	0,95	0,92
0,25	Acurácia	0,60	0,85	0,92	0,92	0,91
	Precisão	1,00	1,00	0,97	0,94	0,92
0,125	Acurácia	0,60	0,85	0,91	0,92	0,91
	Precisão	1,00	1,00	0,96	0,94	0,92
0,0625	Acurácia	0,60	0,84	0,89	0,91	0,90
	Precisão	1,00	1,00	0,95	0,93	0,92
1	Taxa de Acerto	0,56	0,84	0,96	0,98	0,99
	Taxa Falso alarme	0,00	0,00	0,15	0,50	0,75
0,5	Taxa de Acerto	0,55	0,83	0,96	0,98	0,99
	Taxa Falso alarme	0,00	0,00	0,20	0,50	0,75
0,25	Taxa de Acerto	0,56	0,84	0,94	0,98	0,99
	Taxa Falso alarme	0,00	0,00	0,30	0,60	0,75
0,125	Taxa de Acerto	0,55	0,83	0,94	0,98	0,99
	Taxa Falso alarme	0,00	0,00	0,35	0,60	0,80
0,0625	Taxa de Acerto	0,55	0,83	0,93	0,97	0,98
	Taxa Falso alarme	0,00	0,00	0,45	0,60	0,80

O aplicativo determina a seguinte separação exposta na figura 45 ( $\sigma = 1$  e  $\nu = 0,125$ ):

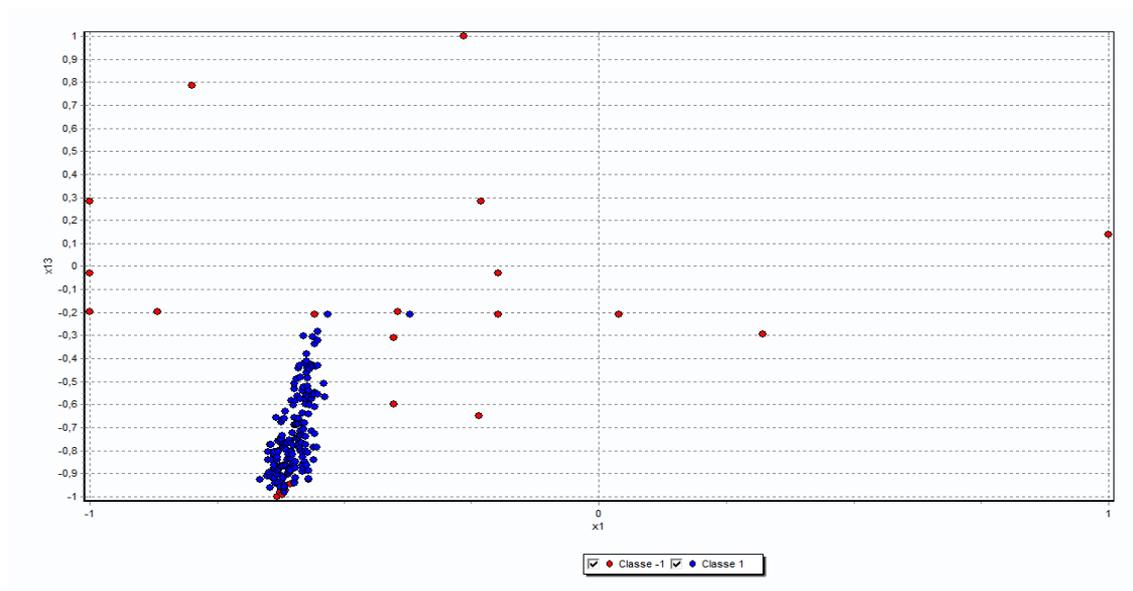


Figura 45 SVM - One Class - Outliers - Reconhecimento de Vinhos

O conjunto de dados Abalone apresenta a seguinte matriz de confusão e precisão expostos nas tabelas 24 e 25 respectivamente:

Tabela 24. SVM - One Class - Matriz de Confusão - Abalone

C	$\nu$	0,5		0,25		0,125	
		CLASSE 1	CLASSE -1	CLASSE 1	CLASSE -1	CLASSE 1	CLASSE -1
1	CLASSE 1	2.103	0	3.155	0	3.681	0
	CLASSE -1	2.074	28	1.022	28	496	28
	Total	4.177	28	4.177	28	4.177	28
0,5	CLASSE 1	2.103	0	3.155	0	3.680	0
	CLASSE -1	2.074	28	1.022	28	497	28
	Total	4.177	28	4.177	28	4.177	28
0,25	CLASSE 1	2.102	0	3.152	0	3.678	0
	CLASSE -1	2.075	28	1.025	28	499	28
	Total	4.177	28	4.177	28	4.177	28
0,125	CLASSE 1	2.100	0	3.153	0	3.680	0
	CLASSE -1	2.077	28	1.024	28	497	28
	Total	4.177	28	4.177	28	4.177	28
0,0625	CLASSE 1	2.104	0	3.153	0	3.680	0
	CLASSE -1	2.073	28	1.024	28	497	28
	Total	4.177	28	4.177	28	4.177	28

C	v	0,0625		0,0313	
		CLASSE 1	CLASSE -1	CLASSE 1	CLASSE -1
1	CLASSE 1	3.941	0	4.075	0
	CLASSE -1	236	28	102	28
	Total	4.177	28	4.177	28
0,5	CLASSE 1	3.941	0	4.072	0
	CLASSE -1	236	28	105	28
	Total	4.177	28	4.177	28
0,25	CLASSE 1	3.944	0	4.074	0
	CLASSE -1	233	28	103	28
	Total	4.177	28	4.177	28
0,125	CLASSE 1	3.942	0	4.072	0
	CLASSE -1	235	28	105	28
	Total	4.177	28	4.177	28
0,0625	CLASSE 1	3.941	0	4.074	0
	CLASSE -1	236	28	103	28
	Total	4.177	28	4.177	28

Tabela 25. SVM - One Class - Outliers - Abalone

c	Índice	v				
		0,5	0,25	0,125	0,0625	0,0313
1	Acurácia	0,51	0,76	0,88	0,94	0,98
	Precisão	1,00	1,00	1,00	1,00	1,00
0,5	Acurácia	0,51	0,76	0,88	0,94	0,98
	Precisão	1,00	1,00	1,00	1,00	1,00
0,25	Acurácia	0,51	0,76	0,88	0,94	0,98
	Precisão	1,00	1,00	1,00	1,00	1,00
0,125	Acurácia	0,51	0,76	0,88	0,94	0,98
	Precisão	1,00	1,00	1,00	1,00	1,00
0,0625	Acurácia	0,51	0,76	0,88	0,94	0,98
	Precisão	1,00	1,00	1,00	1,00	1,00
1	Taxa de Acerto	0,50	0,76	0,88	0,94	0,98
	Taxa Falso alarme	0,00	0,00	0,00	0,00	0,00
0,5	Taxa de Acerto	0,50	0,76	0,88	0,94	0,97
	Taxa Falso alarme	0,00	0,00	0,00	0,00	0,00
0,25	Taxa de Acerto	0,50	0,75	0,88	0,94	0,98
	Taxa Falso alarme	0,00	0,00	0,00	0,00	0,00
0,125	Taxa de Acerto	0,50	0,75	0,88	0,94	0,97
	Taxa Falso alarme	0,00	0,00	0,00	0,00	0,00
0,0625	Taxa de Acerto	0,50	0,75	0,88	0,94	0,98
	Taxa Falso alarme	0,00	0,00	0,00	0,00	0,00

O aplicativo expõe a separação dos dados anômalos conforme figura 46 abaixo:

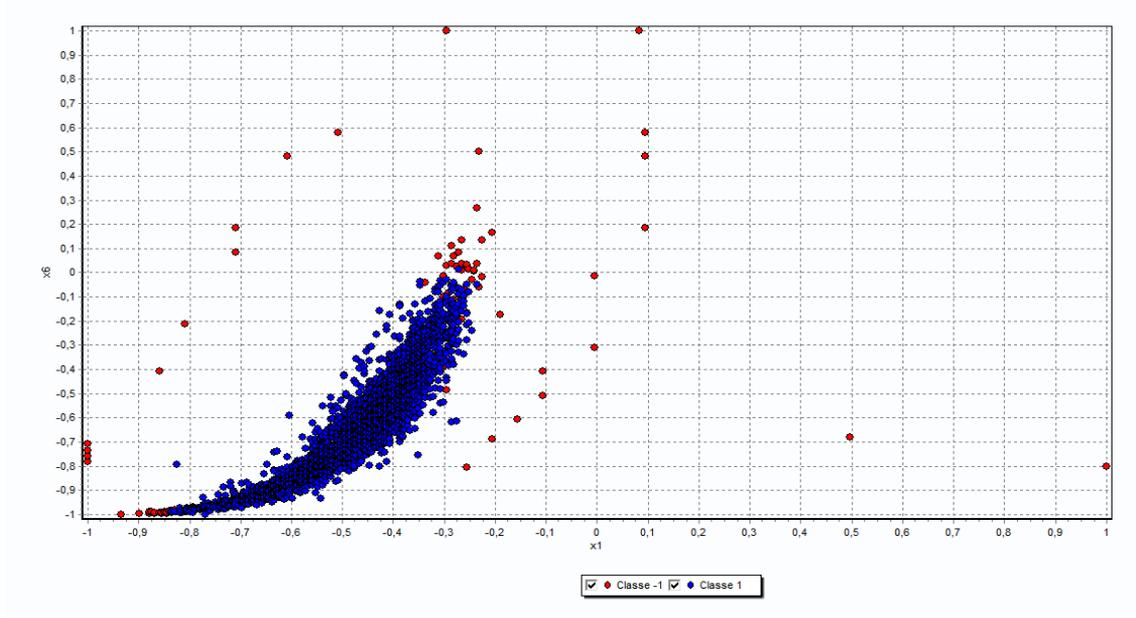


Figura 46 SVM - One Class - Outliers – Abalone

#### 4.4.3 Precisão do aplicativo Função de Similaridade Média

Faz-se a mesma ressalva, antes da submissão dos dados artificiais a este algoritmo, a respeito dos valores de  $\sigma$  a serem aplicados à função Kernel, face a diferença de implementação entre o aplicativo LIBSVM e este, conforme exposto em 4.4.1.

A submissão dos dados da Planta Íris com *outliers* a este algoritmo demonstra que é capaz de elencar os dados anômalos, resultando na matriz de confusão da tabela 26 e precisão exposta na tabela 27:

Tabela 26. FSM - Matriz de Confusão - Planta Íris

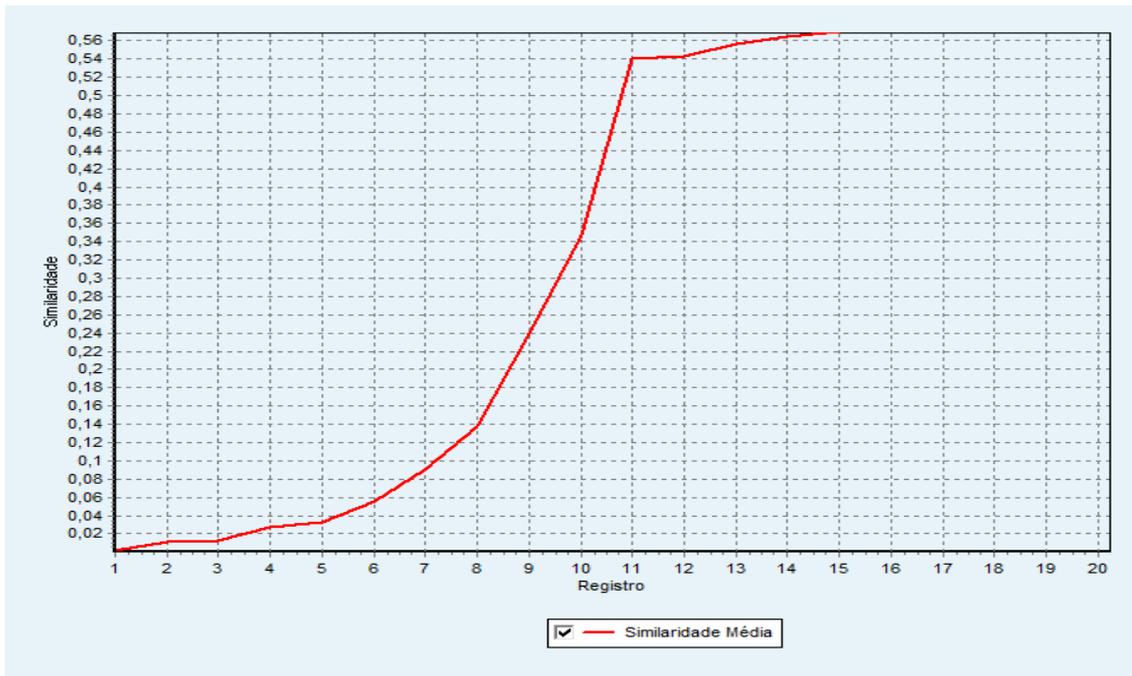
$\sigma$	0,25		0,5		1	
	Normal	Anômalo	Normal	Anômalo	Normal	Anômalo
Normal	150	0	150	0	150	0
Anômalo	0	10	0	10	0	10
Total	150	10	150	10	150	10

$\sigma$	2		4	
	Normal	Anômalo	Normal	Anômalo
Normal	150	0	150	0
Anômalo	0	10	0	10
Total	150	10	150	10

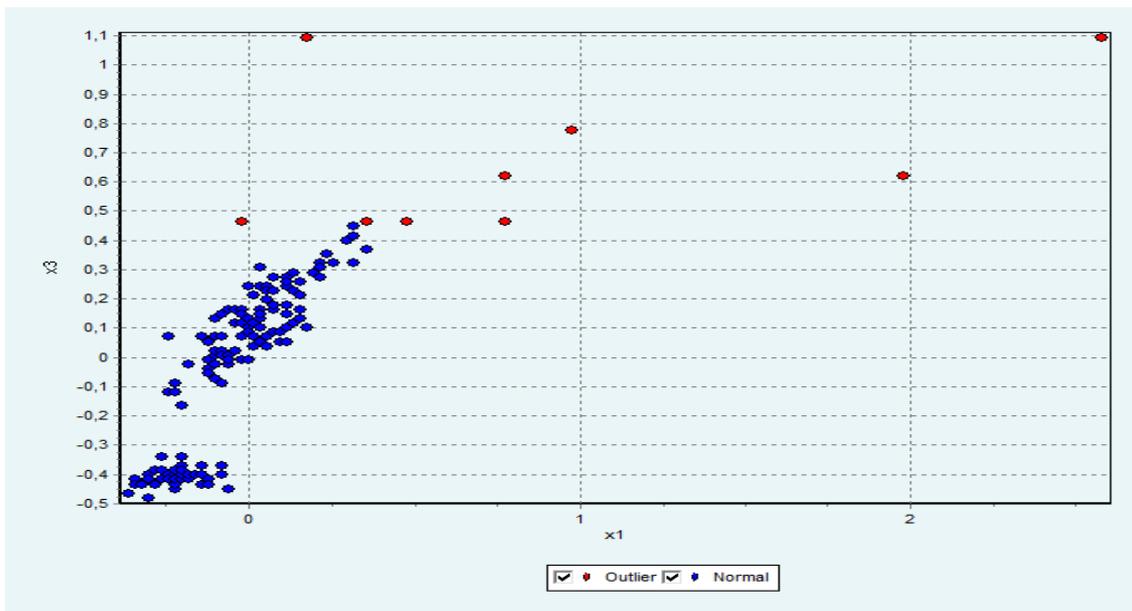
**Tabela 27.** FSM - Índices de precisão - Planta Íris

Índice	$\sigma$				
	0,25	0,5	1	2	4
Acurácia	1,00	1,00	1,00	1,00	1,00
Precisão	1,00	1,00	1,00	1,00	1,00
Taxa de Acerto	1,00	1,00	1,00	1,00	1,00
Taxa Falso alarme	0,00	0,00	0,00	0,00	0,00

Tal algoritmo evidencia a curva de similaridade e exposição dos *outliers* expostos nas figuras 47 e 48 respectivamente:



**Figura 47** FSM - Curva de Similaridade - Planta Íris



**Figura 48** FSM - Outliers - Planta Íris

Submetendo-se a base de dados de Reconhecimento de Vinhos, verifica-se a eficácia do algoritmo, que evidencia a matriz de confusão e índices de precisão expostos nas tabelas 28 e 29 respectivamente:

**Tabela 28.** FSM - Matriz de Confusão - Reconhecimento de Vinhos

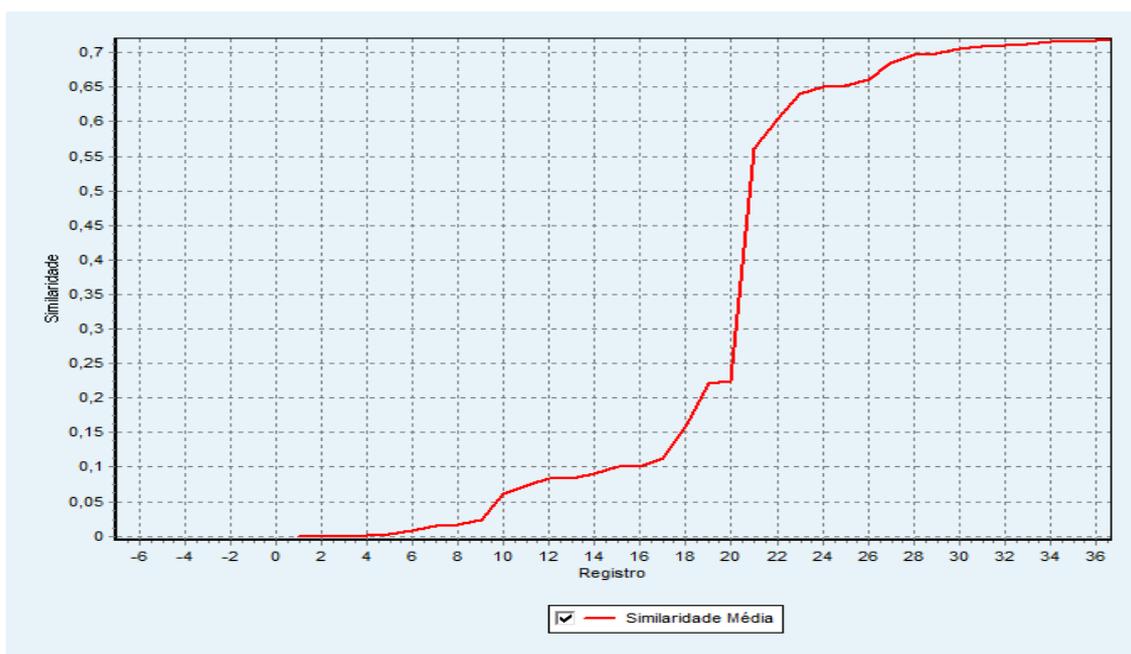
$\sigma$	0,25		0,5		1	
	Normal	Anômalo	Normal	Anômalo	Normal	Anômalo
Normal	178	0	178	0	178	0
Anômalo	0	20	0	20	0	20
Total	178	20	178	20	178	20

$\sigma$	2		4	
	Normal	Anômalo	Normal	Anômalo
Normal	178	0	178	0
Anômalo	0	20	0	20
Total	178	20	178	20

**Tabela 29.** FSM - Índices de precisão - Reconhecimento de Vinhos

Índice	$\sigma$				
	0,25	0,5	1	2	4
Acurácia	1,00	1,00	1,00	1,00	1,00
Precisão	1,00	1,00	1,00	1,00	1,00
Taxa de Acerto	1,00	1,00	1,00	1,00	1,00
Taxa Falso alarme	0,00	0,00	0,00	0,00	0,00

O aplicativo determina a seguinte curva de Similaridade da figura 49e evidencia os *outliers* expostos na figura 50:



**Figura 49** FSM - Curva de Similaridade - Reconhecimento de Vinhos

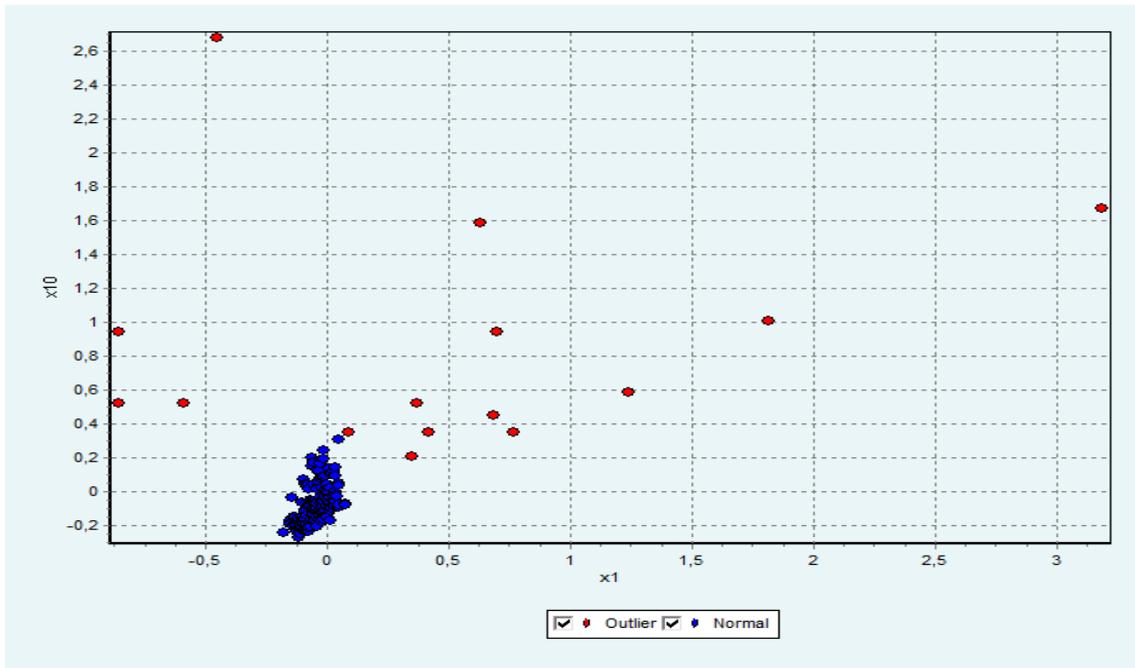


Figura 50 FSM - Outliers - Reconhecimento de Vinhos

Aplicando-se o algoritmo à base de dados do molusco Abalone obtém-se a seguinte matriz de confusão da tabela 30 e precisão exposta na tabela 31:

Tabela 30. FSM - Matriz de Confusão - Abalones

$\sigma$	0,25		0,5		1	
	Normal	Anômalo	Normal	Anômalo	Normal	Anômalo
Normal	4.177	0	4.177	0	4.177	0
Anômalo	0	28	0	28	0	28
Total	4.177	28	4.177	28	4.177	28

$\sigma$	2		4	
	Normal	Anômalo	Normal	Anômalo
Normal	4.177	0	4.177	0
Anômalo	0	28	0	28
Total	4.177	28	4.177	28

Tabela 31. FSM - Índices de Precisão - Abalones

Índice	$\sigma$				
	0,25	0,5	1	2	4
Acurácia	1,00	1,00	1,00	1,00	1,00
Precisão	1,00	1,00	1,00	1,00	1,00
Taxa de Acerto	1,00	1,00	1,00	1,00	1,00
Taxa Falso alarme	0,00	0,00	0,00	0,00	0,00

O aplicativo expõe a curva de Similaridade da figura 51 e evidencia os *outliers*, conforme figura 52:

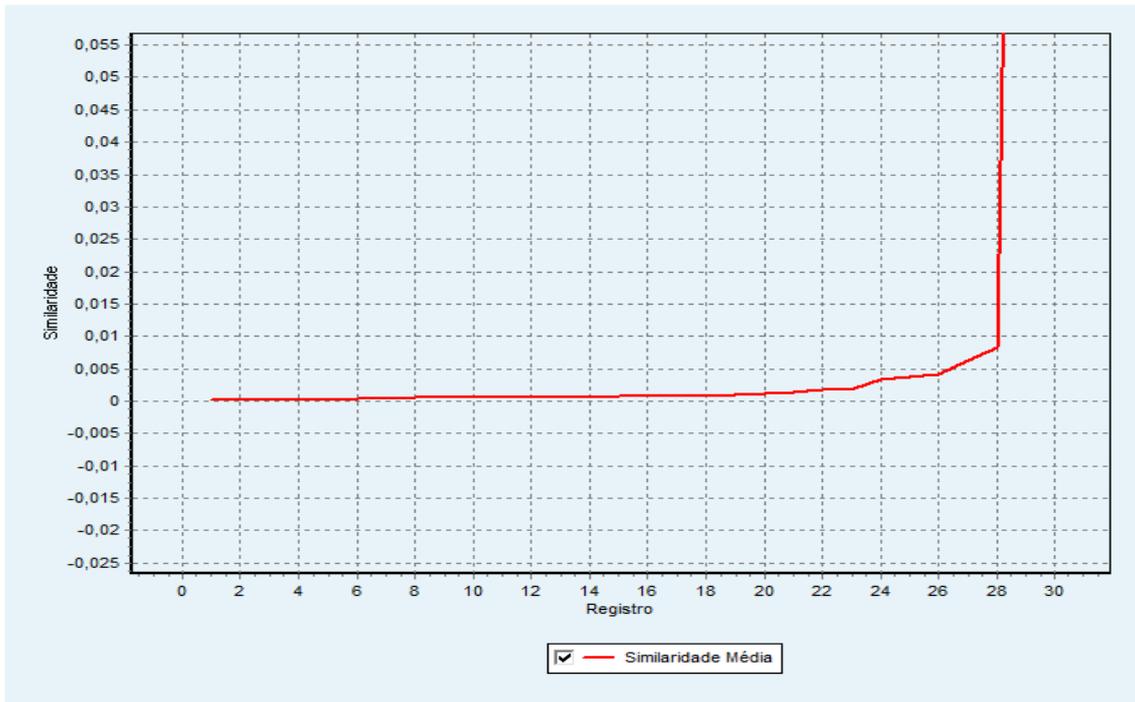


Figura 51 FSM - Curva de Similaridade - Abalones

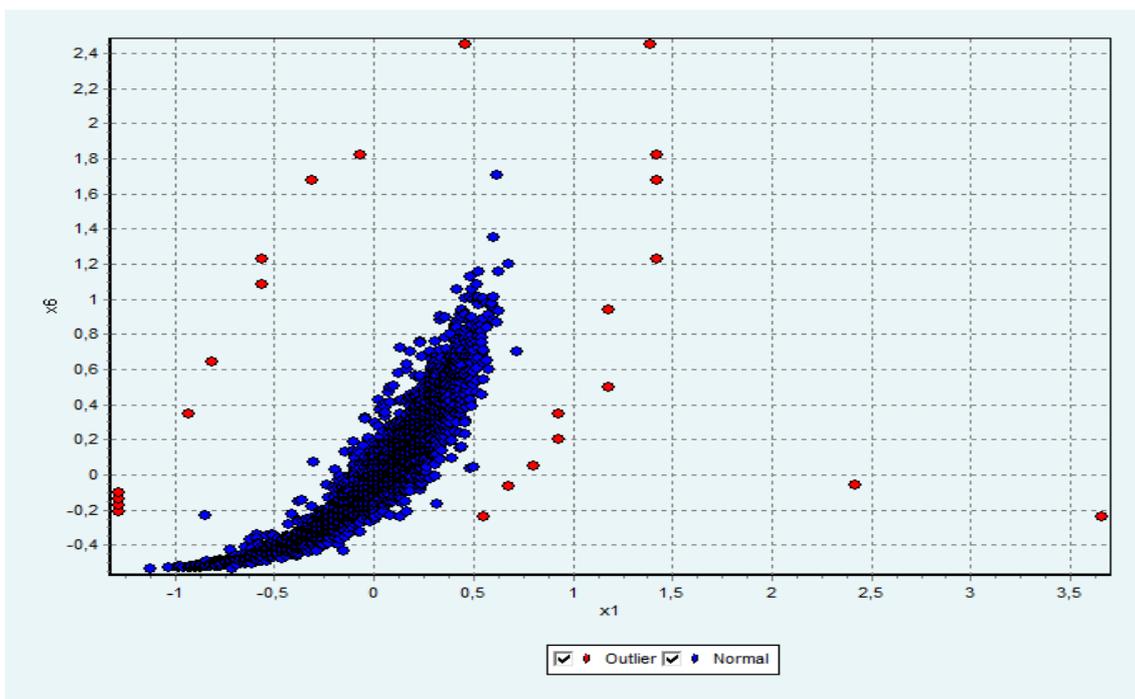


Figura 52 FSM - Outliers - Abalones

#### 4.5 Comparação com outros resultados

Conforme expõe AL-ZOUBI, M. B. (2009), a base de dados da Planta Íris contém 10 *outliers* e há 48 dados anômalos na base de *Liver Disorders* expostos na tabela 32. Estas duas base estão disponíveis em UCI Machine Learning Repository (<http://www.ics.uci.edu/~mlern/MLRepository.html>).

**Tabela 32.** Registros Anômalos

Base de Dados	Número dos registros anômalos
Planta Iris	106; 107;108; 110; 118; 119; 123; 131; 132; 136
<i>Liver Disorder</i>	2; 20; 22; 25; 36; 77; 85; 111; 115; 123; 133; 134; 139; 148; 157; 167; 168; 175; 179; 182; 183; 186; 187; 189; 190; 205; 224; 233; 252; 261; 278; 286; 294; 300; 307; 311; 312; 313; 316; 317; 323; 326; 331; 335; 337; 342; 343; 345

O autor propõe uma variação do algoritmo *Partitioning Around Medoids* (PAM) que consegue identificar 8 outliers na primeira base e 39, na segunda base.

SHAARI et al. (2007), ao buscar uma nova técnica de evidenciar outliers (denominada RSetOF), submeteram a base de dados *Breast Cancer* obtida no mesmo sítio acima, que contém 458 registros classificados como benignos e 241, malignos. Os autores informam que para formar uma base desbalanceada, removeram alguns registros ditos malignos, retendo um em cada seis e retirando dos registros benignos aqueles com valores faltantes. A base final ficou contendo 444 registros benignos e 39 malignos, sendo, então, submetida ao algoritmo proposto pelos autores, que conseguiu elencar todos os registros malignos, entendidos como *outliers*. Apesar dos autores não especificarem quais registros constituíam a base, pode-se reproduzir a base final à luz da forma descrita pelos mesmos.

Aplicando os algoritmos de Agrupamento Nebuloso (KPCM), *SVM – One Class* (implementado no LIBSVM) e Função de Similaridade Média (FSM) a estas bases, obtém-se os seguintes resultados da tabela 33, atribuindo-se os valores  $c = 0,5$  e  $\nu = 0,25$  para os parâmetros do algoritmo *SVM – One Class*, e  $\sigma = 1$  para os algoritmos KPCM e FSM.

**Tabela 33.** Quantidade de *outliers* encontrados

Algoritmo		KPCM	SVM – One Class	FSM
Base				
N. Outliers corretamente identificados	Planta Iris	8	6	8
	<i>Liver Disorder</i>	41	36	41
	<i>Breast Cancer</i>	39	36	39

O resultado demonstra que os algoritmos KPCM, *SVM – One Class* e FSM conseguem elencar a maioria dos *outliers*, tendo praticamente o mesmo resultado acusado pelo PAM e RSetOF.

AL-ZOUBI, M. B. (2009) e SHAARI et al. (2007) não apresentam índices de precisão algum, limitando-se a informar a quantidade de *outliers* evidenciados. Para

os algoritmos analisados neste estudo, pode-se evidenciar a precisão alcançada, conforme exposto na tabela 34:

**Tabela 34.** Precisão dos algoritmos

Base de Dados	Índice	Algoritmo		
		KPCM	SVM – One Class	FSM
Planta Iris	Acurácia	0,85	0,88	0,89
	Precisão	0,98	0,97	0,98
	Taxa de Acerto	0,85	0,90	0,89
	Taxa Falso alarme	0,20	0,40	0,20
Liver Disorder	Acurácia	0,91	0,82	0,90
	Precisão	0,97	0,95	0,97
	Taxa de Acerto	0,92	0,83	0,91
	Taxa Falso alarme	0,15	0,25	0,15
Breast Cancer	Acurácia	1,00	0,94	1,00
	Precisão	1,00	0,99	1,00
	Taxa de Acerto	1,00	0,94	1,00
	Taxa Falso alarme	0,00	0,08	0,00

Este último teste confirma não trivialidade da evidenciação dos outliers, mormente quando os valores não são extremos, mas entremeados no conjunto. Ainda assim, verifica-se que os algoritmos propostos conseguem estabelecer os dados anômalos com boa precisão, ainda mais, quando comparados com os resultados estabelecidos por CAMASTRA, F. e VERRI, A. (2005) ao testarem a base Íris com diversos algoritmos, que registraram uma precisão entre 0,81 e 0,94.

#### 4.6 Considerações

Dos resultados desses algoritmos, pode-se aquilatar que todos são eficazes e permitem algumas considerações baseadas na precisão alcançada:

- A melhor acurácia, precisão e taxa de falso alarme para o aplicativo LIBSVM são encontradas quando o parâmetro  $\nu$  assume valores entre 0,25 e 0,06125.
- Os algoritmos apresentam uma precisão adequada, cujo valor pode sofrer variação de acordo com o parâmetro utilizado no algoritmo (mormente para o SVM-ONE CLASS), e pode ser observada no gráfico ROC (*Receiver Operating Characteristic*) abaixo (Figura 13). Neste gráfico os valores da taxa de falsos alarmes são inscritos no eixo das abcissas e aqueles da taxa de verdadeiros positivos no eixo das ordenadas, sendo que a melhor precisão é alcançada quanto mais próximo se está da coordenada (0,1),



## 5. Descrição da Base de Dados

A possibilidade de extrair conhecimento dos registros relativos à saúde foi propugnada por diversos pesquisadores, que entenderam que os dados têm a finalidade de viabilizar a administração e a fiscalização do cumprimento de obrigações. Estes observam que há disponibilidade de grandes massas de dados, mas sem a consequente extração de conhecimento, uma vez que a ótica tem sido meramente administrativa, no sentido de registrar o fato. Vêem que este acervo (vide figura 54) deve apoiar a governança do setor e prover suporte à decisão (VASCONCELOS, M.M *et al.*, 2002).

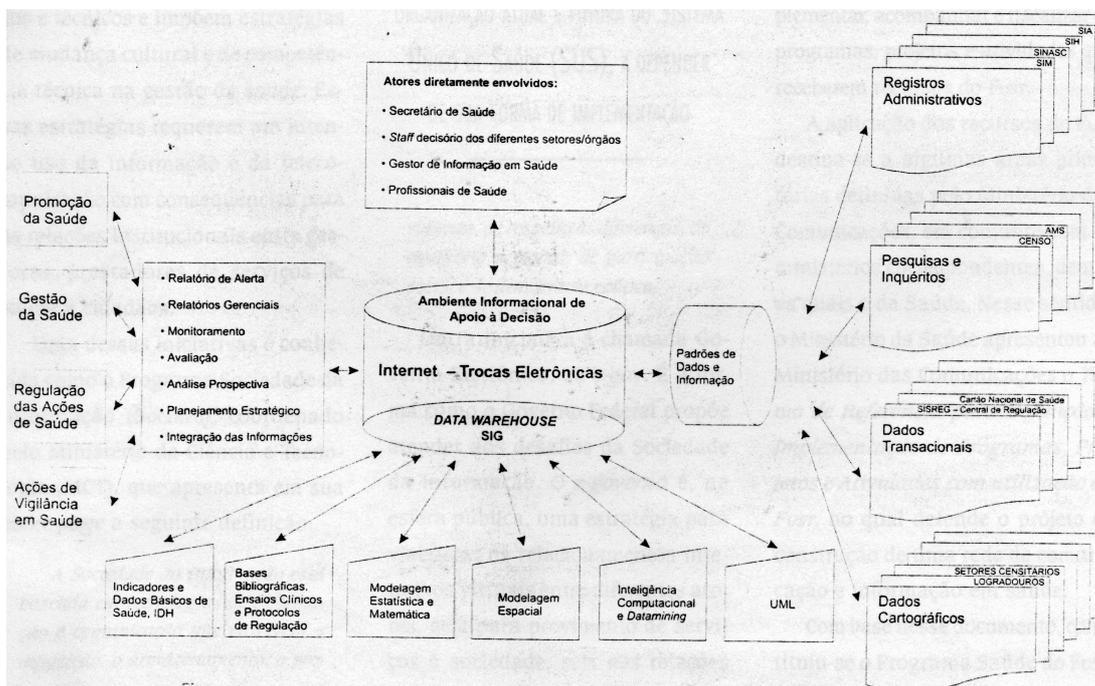


Figura 54. Possibilidades de suporte à descoberta do conhecimento (Vasconcelos, M.M *et al.*, 2002).

Neste sentido, cinge-se o foco deste trabalho nos dados constituintes do Movimento de Autorização de Internação Hospitalar (AIH), limitados ao estado do Rio de Janeiro. Tal base contempla os lançamentos de todos os procedimentos executados nos hospitais públicos e constitui um universo ímpar para utilização de técnicas de mineração.

### 5.1 Atributos da Base de Dados.

O sistema de controle da internação hospitalar apresenta as seguintes informações básicas, cujos campos constituintes da base estão listados no anexo A:

- Caracterização do paciente (idade, sexo, residência), da internação (número da AIH, hospital, especialidade, procedimento solicitado e realizado, diagnóstico, data de internação e alta, motivo de cobrança) e de faturamento (valores cobrados).
- Ocorrências valoradas de cada ato médico realizado, relativo a serviços profissionais (SP) e serviços auxiliares de diagnose e terapia (SADT), identificando o prestador e o tipo de ato.
- Os Procedimentos autorizados.
- As ocorrências de órtese e prótese de cada AIH, identificando o prestador e o tipo de ato médico.
- Os valores processados por hospital, discriminando-os por faturamentos (normais e complementares), por ordens de recebimento, por adiantamentos e por pagamentos suplementares de competências passadas (incluindo os fatores de recomposição).
- Os valores pagos a cada prestador (terceiro).

Ressalta-se a opção por centrar a análise sobre os dados financeiros, uma vez que, normalmente, estes são o foco de uma auditoria. A base é constituída por 30.000 registros referentes aos lançamentos de dois anos referentes às entidades sediadas no estado do Rio de Janeiro.

## 5.2 Análise Exploratória dos Dados

A tarefa de exame dos dados é uma etapa importante no entendimento dos dados e suas relações. Ainda que o objetivo deste estudo seja o agrupamento dos dados e seus elementos anômalos, a caracterização estatística é útil e auxilia o entendimento deste problema. Neste sentido, dispõe-se da análise exploratória de dados, que busca sintetizar o entendimento por meio de tabelas e gráficos, e da estatística descritiva, que permite verificar medidas de tendência, distribuição dos dados e a existência de *outliers*.

Assim, obtém-se da base de dados a seguinte síntese:

- a) Estatísticas básicas - As medidas descritivas, tais como média, mediana, quartis e o desvio padrão fornecem um resumo das características em relação à tendência central e da dispersão dos dados, consistindo nos seguintes valores (tabela 35) :

**Tabela 35.** Medidas descritivas dos dados

Variável	Máximo	Média	Desvio padrão	Mediana	1 Quartil	2 Quartil	3 Quartil
VAL_SH	8637,03	265,67	305,79	180,72	95,30	180,72	343,03
VAL_SP	2141,39	71,75	65,89	46,62	29,66	46,62	111,43
VAL_SADT	1540,12	14,74	35,53	5,38	1,15	5,38	16,17
VAL_RN	29,76	2,49	5,59	0,00	0,00	0,00	0,00
VAL_ORTP	8670,18	13,91	177,11	0,00	0,00	0,00	0,00
VAL_SANGUE	1067,70	1,57	16,81	0,00	0,00	0,00	0,00
VAL_SADTSR	0,00	0,00	0,00	0,00	0,00	0,00	0,00
VAL_TRANSP	2160,00	0,22	19,55	0,00	0,00	0,00	0,00
VAL_TOT	13160,47	370,36	467,24	275,55	193,42	275,55	422,11
VAL_UTI	8407,50	17,73	153,52	0,00	0,00	0,00	0,00
US_TOT	7477,54	210,43	265,48	156,56	109,90	156,56	239,84

- b) Aderência à Função de Distribuição de Probabilidade Padronizada – aplicando-se o teste de Kolmogorov-Smirnov, observa-se que os valores deste teste são superiores aos limites para os dados considerados, não permitindo afirmar que os dados apresentam comportamento Normal ou Exponencial, conforme pode ser visto nas tabelas 36 e 37. Este teste visa verificar a hipótese de que os dados da amostra em análise são oriundos de uma população com uma determinada distribuição. É baseado na verificação da maior diferença entre a frequência acumulada observada e a estimada pela distribuição pretendida (no caso, normal ou exponencial) com um nível de significância escolhido, normalmente 95% ( $1-\alpha$ ). A decisão por aceitar ou rejeitar a hipótese dos dados serem aderentes a uma determinada distribuição é delimitada pelo valor crítico (p-valor), se menor que 1%, indica que os dados não obedecem à função testada.

**Tabela 36.** Teste Komolgorov-Smirnov (Curva Normal)

Variável	Média	Desvio Padrão	Maior diferença absoluta	Kolmogorov-Smirnov (z)	Nível Crítico
VAL_SH	265,67	305,79	0,21	36,72	0,00
VAL_SP	71,75	65,89	0,19	33,13	0,00
VAL_SADT	14,74	35,53	0,34	58,73	0,00
VAL_RN	2,49	5,59	0,51	87,55	0,00
VAL_ORTP	13,91	177,11	0,51	87,63	0,00
VAL_SANGUE	1,57	16,81	0,51	88,80	0,00
VAL_TRANSP	0,22	19,55	0,50	87,36	0,00
VAL_TOT	370,36	467,24	0,24	41,23	0,00
VAL_UTI	17,73	153,52	0,51	89,10	0,00
US_TOT	210,43	265,48	0,24	41,23	0,00

**Tabela 37.** Teste Kolmogorov - Smirnov (Curva Exponencial)

Variável	Média	Maior diferença absoluta	Kolmogorov-Smirnov (z)	Nível Crítico
VAL_SH	265,8336	0,157	27,194	0,00
VAL_SP	71,7972	0,161	27,899	0,00
VAL_SADT	16,8719	0,348	56,395	0,00
VAL_RN	14,9574	5,366	379,366	0,00
VAL_ORTP	549,02	38,822	1070,255	0,00
VAL_SANGUE	63,9218	39,77	1081,121	0,00
VAL_TRANSP	1336,38	5998,999	13414,17	0,00
VAL_TOT	370,5837	0,197	34,133	0,00
VAL_UTI	562,1169	30,828	948,18	0,00
US_TOT	210,5587	0,197	34,134	0,00

- c) Existência de outliers - Considerando o limite proposto no gráfico Box-plot, onde todo valor superior a  $3^{\circ}$  quartil +  $1,5 * (3^{\circ}$  quartil -  $1^{\circ}$  quartil) é considerado um outlier, tem-se as seguintes quantidade de outliers, segundo cada variável (tabela 38):

**Tabela 38.** Quartis e outliers

Variável	1 Quartil	3 Quartil	Limite Superior	Quantidade de Outliers	Percentual em relação ao total
VAL_SH	95,30	343,03	714,625	1065	3,55
VAL_SP	29,66	111,43	234,085	516	1,72
VAL_SADT	1,15	16,17	38,7	2133	7,11
VAL_RN	0,00	0,00	0	4999	16,66
VAL_ORTP	0,00	0,00	0	760	2,53
VAL_SANGUE	0,00	0,00	0	739	2,46
VAL_TRANSP	0,00	0,00	0	5	0,02
VAL_TOT	193,42	422,11	765,145	1657	5,52
VAL_UTI	0,00	0,00	0	946	3,15
US_TOT	109,90	239,84	434,75	1657	5,52

Utilizando-se o método proposto por SCHWERTMAN, N. C. *et al.* (2004) para o estabelecimento de *outliers*, obtém-se os valores expostos na tabela 39:

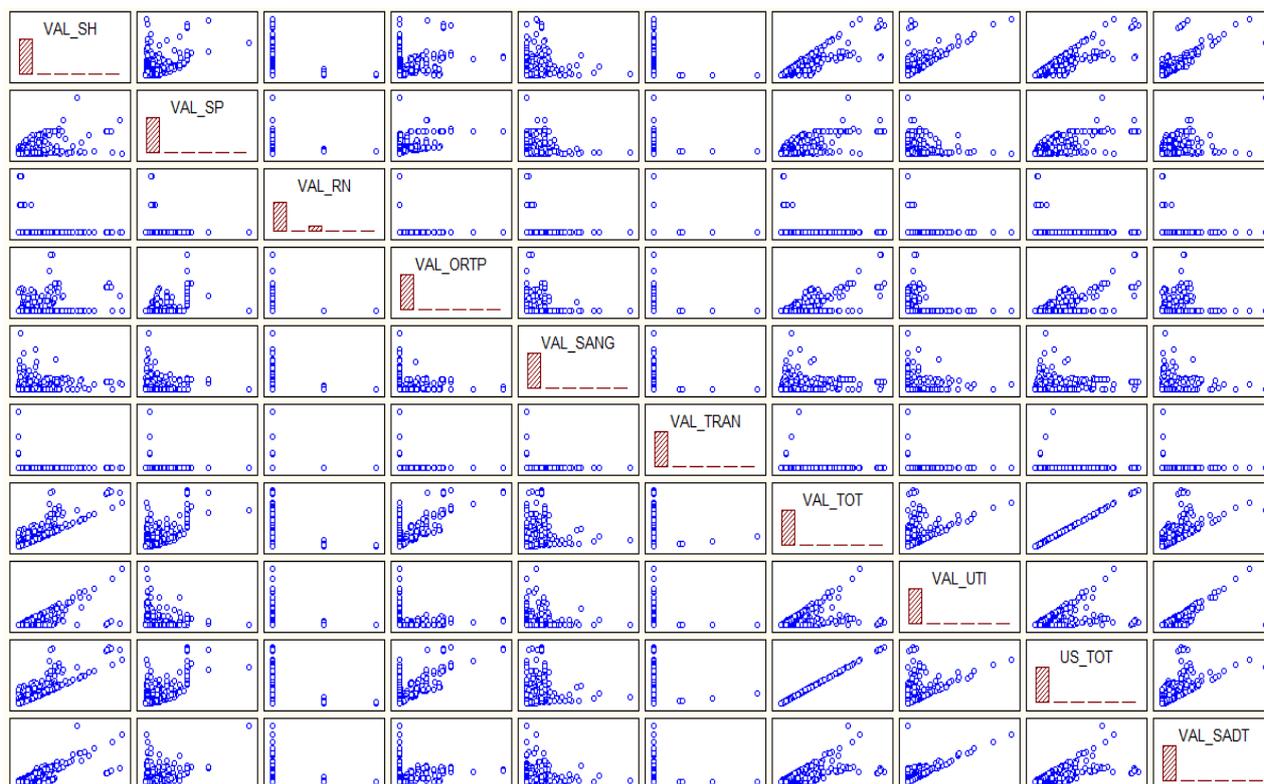
**Tabela 39** - Quantidade de outliers - SCHWERTMAN, N. C. *et al.* (2004)

Variável	1 Quartil	2 Quartil	3 Quartil	Limite Superior	Quantidade de Outliers	Percentual em relação ao total
VAL_SH	95,30	180,72	343,03	576,5746	3396	11,32
VAL_SP	29,66	46,62	111,43	204,6838	803	2,68
VAL_SADT	1,15	5,38	16,17	31,69551	2822	9,41
VAL_RN	0,00	0	0,00	0	4999	16,66
VAL_ORTP	0,00	0	0,00	0	760	2,53
VAL_SANGUE	0,00	0	0,00	0	739	2,46
VAL_TRANSP	0,00	0	0,00	0	5	0,02
VAL_TOT	193,42	275,55	422,11	632,9922	3833	12,78
VAL_UTI	0,00	0	0,00	0	946	3,15
US_TOT	109,90	156,56	239,84	359,6699	3833	12,78

- d) Matriz de Correlação, que apresenta a medida de relação entre as variáveis, quantificando a dependência entre as dimensões, considerando-se significativas aquelas superiores a 0,5, sendo que quanto mais próximo de 1, mais forte é a correlação (SANTOS, J. S. *et al.*, 2008) (HIROTA, K.; PEDRICZ, W., 1999). Obtiveram-se as seguintes correlações expostas na tabela 40 e nos gráficos de dispersão da figura 55:

**Tabela 40.** Matriz de correlação

	VAL_SH	VAL_SP	VAL_RN	VAL_ORTP	VAL_SANGUE	VAL_TRANSP	VAL_TOT	VAL_UTI	US_TOT	VAL_SADT
VAL_SH	1,00									
VAL_SP	0,34	1,00								
VAL_RN	-0,19	0,27	1,00							
VAL_ORTP	0,35	0,44	-0,04	1,00						
VAL_SANGUE	0,28	0,17	-0,04	0,17	1,00					
VAL_TRANSP	-0,01	0,00	-0,01	0,00	0,00	1,00				
VAL_TOT	0,90	0,57	-0,10	0,70	0,33	0,04	1,00			
VAL_UTI	0,62	0,12	-0,05	0,12	0,24	0,00	0,54	1,00		
US_TOT	0,90	0,57	-0,10	0,70	0,33	0,04	1,00	0,54	1,00	
VAL_SADT	0,71	0,40	-0,16	0,32	0,32	0,00	0,73	0,79	0,73	1,00



**Figura 55** Gráficos de dispersão

Vê-se que a variável US\_TOT (valores pagos em dólares) é fortemente correlacionada com a variável VAL\_TOT (Valor total da AIH) e poderia ser explicada por esta última. Apesar de não apresentar uma correlação acentuada, pode-se verificar um acoplamento entre as variáveis VAL\_TOT e VAL\_SH (Valor de serviços hospitalares).

e) Análise de componentes principais (PCA) – o excesso de variáveis analisadas simultaneamente pode dificultar a interpretação de resultados, uma vez que o ser humano tem facilidade de perceber apenas três dimensões. A PCA determina a importância de cada variável, de forma a reduzir a dimensão para um número administrável (PEREIRA, E. B. B; PEREIRA, M. B., 2004) (BUHAGIAR, A., 2002) (NETO, J. M. M.; MOITA, G. C., 1998) (SCHMITT, J., 2005).

Para estabelecer os componentes principais há vários critérios, podendo-se optar:

- critério de Kaiser, onde os componentes principais a serem considerados são aqueles com autovalor igual ou superior a um

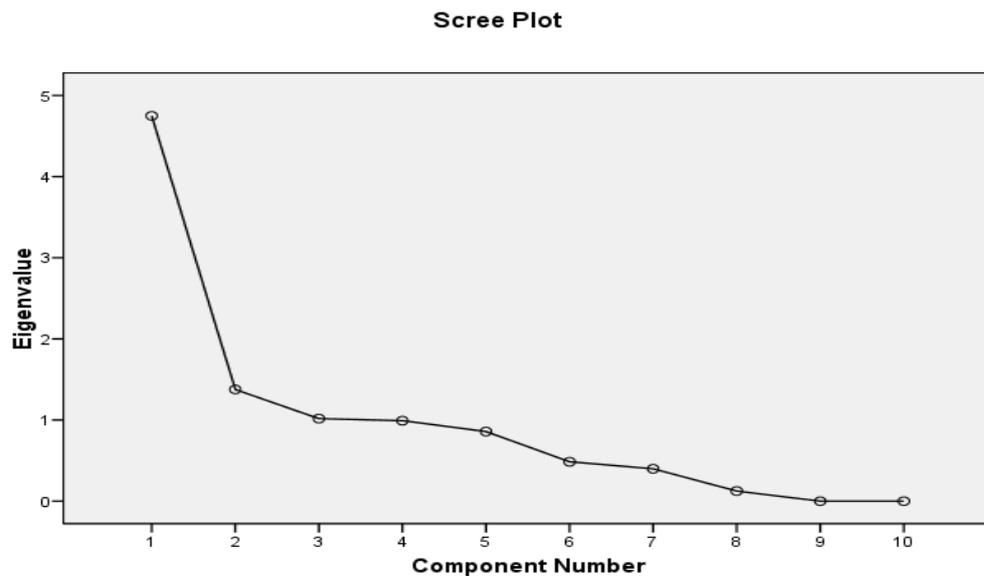
e, neste caso, resume-se aos três primeiros componentes (observe-se a tabela 41);

- critério baseado na percentagem acumulada da variância explicada, onde se considera suficientes as componentes que explicam mais de 70% da variância explicada acumulada, determinando-se, assim, que os três primeiros componentes caracterizam os dados (vide tabela 41);

**Tabela 41.** Componentes Principais

Componente	Autovalor		
	Total	Variância Explicada (%)	Variância Explicada Acumulada (%)
1	4,75	47,49	47,49
2	1,38	13,76	61,25
3	1,02	10,17	71,42
4	0,99	9,92	81,34
5	0,86	8,58	89,92
6	0,49	4,85	94,77
7	0,40	3,98	98,76
8	0,12	1,24	100,00
9	0,00	0,00	100,00
10	0,00	0,00	100,00

- Critério do Scree plot, cujo gráfico representa a percentagem de variância explicada por cada componente e é uma forma gráfica representativa dos dados da tabela 40. Quando a percentagem se reduz e a curva passa a ser quase paralela às abcissas, determina-se que as componentes à esquerda são as componentes principais, como pode ser observado na figura 57 abaixo, onde, novamente, vê-se a primazia dos três primeiros componentes.



**Figura 56.** Scree Plot

Pode-se observar na tabela 42 a contribuição de cada componente às dimensões em análise, isto é, quais dimensões são explicadas por cada componente.

Conforme exposto por BUHAGIAR, A. (2002), a rotação dos componentes consegue simplificar os fatores, tornando-os maiores ou menores e os retirando de uma situação intermediária, bem como evita destes ficarem restritos a um só componente. O resultado desta operação está exposto na tabela 42. HAIR, J. F *et al.* (2007) expõem que a rotação permite atingir um padrão fatorial mais simples e significativo, explicitando a contribuição de cada variável (quanto maior a carga, mais significativa). Dessa forma, pode-se observar maior relevância das variáveis VAL\_SH (valor dos serviços hospitalares), VAL\_SADT (valor dos serviços auxiliares de diagnóstico e terapia) e VAL\_TOT (valor total da Autorização de Internação Hospitalar).

Tabela 42. Carga dos fatores

Matriz de Componentes				Matriz de componentes rotacionados		
Dimensão	Componente			Componente		
	1	2	3	1	2	3
VAL_SH	0,885	-0,211	0,008	0,814	0,375	-0,153
VAL_SP	0,574	0,628	0,095	0,168	0,773	0,328
VAL_SADT	0,844	-0,283	0,175	0,874	0,244	-0,035
VAL_RN	-0,11	0,653	0,543	-0,243	0,234	0,787
VAL_ORTP	0,627	0,44	-0,349	0,161	0,812	-0,155
VAL_SANGUE	0,407	-0,078	0,172	0,42	0,132	0,087
VAL_TRANSP	0,017	0,022	-0,621	-0,2	0,229	-0,543
VAL_TOT	0,976	0,102	-0,114	0,688	0,698	-0,125
VAL_UTI	0,67	-0,458	0,346	0,88	-0,045	0,05
US_TOT	0,976	0,102	-0,114	0,688	0,698	-0,125

## 6. Resultados

Viu-se que os métodos propostos para evidenciação dos outliers são eficazes. Assim, após testes sobre dados conhecidos, passa-se a aplicá-los sobre a base de informações hospitalares.

Para todos os resultados dos algoritmos, determina-se as dimensões VAL\_SH (valor de serviços hospitalares) e VAL\_TOT (valor total da Autorização de Internação Hospitalar) como as representativas para confecção do gráfico face aos valores altos assinalados a estas na análise de fatores rotacionados, bem como o significado destes atributos na caracterização do gasto. Ressalta-se que se busca com os gráficos apenas a visualização de que os algoritmos têm uma concordância quantos aos registros cancelados como *outliers*. Entretanto, o estabelecimento dessa concordância dar-se-á por meio do confronto de resultados, onde o gráfico indica esta tendência.

### 6.1 Agrupamento Nebuloso (KPCM)

Análise de agrupamento é uma ferramenta útil para extração de conhecimento em uma base de dados, cujo objetivo é ajuntar dados semelhantes em  $k$  grupos. A medida de semelhança é baseada na distância entre um registro em particular e o centro de um dado *cluster*.

Neste estudo, o objetivo, ainda que baseado nesta metodologia, não busca estabelecer os grupos de dados, mas evidenciar os dados que não pertencem a grupo algum.

A análise clássica de agrupamentos, neste caso, não é útil, pois determina que cada elemento tem de pertencer exatamente a um grupo de dados, não permitindo a existência de elementos anômalos. Entretanto, pode-se relaxar tal regra, dando origem a um agrupamento nebuloso, onde todos os dados pertencem parcialmente a todos os *clusters*. Como visto, os elementos distantes de todos os grupos terão pertinência a todos os *clusters* baixos, próximo de zero e, desta forma, pode-se caracterizar um elemento anômalo por intermédio dessa característica, isto é, considerar-se-á como tal todo elemento cuja soma dos graus de pertencimento aos grupos seja próxima a zero.

Neste mister, utilizar-se-á o algoritmo *Kernel Possibilistic C-Means* (KPCM), analisando-se os resultados, quando se varia o número de *clusters* e o parâmetro  $\sigma$  da função kernel para a determinação de distâncias no espaço característico.

Assumindo que toda soma das pertinências iguais ou inferiores a 0,00001 (considera-se que este valor está suficientemente próximo de zero) caracteriza o dado como anômalo, chega-se às seguintes quantidades de *outliers* expostas na tabela 43:

Tabela 43. Quantidade de dados normais e anômalos (KPCM)

N.Cluster	Parâmetro $\sigma$				
	0,25	0,5	1	2	4
2	1873	2395	2611	2597	2388
3	0	1184	1406	1440	1637
4	889	1739	1764	1717	1695
5	2416	2266	2055	1856	1739
6	0	0	514	761	841
7	2563	2452	2259	2072	1864

Quando do agrupamento em três grupos e  $\sigma = 0,25$ , vê-se que não há registros com pertinências totais inferiores ao limite proposto (observe-se figura 57), mas aumentando-se (empiricamente, partindo-se de 0,00001) o limite verifica-se que, ao se alcançar o valor 0,00007, ocorrem 638 dados anômalos. Os dados anômalos estão representados na figura 58.

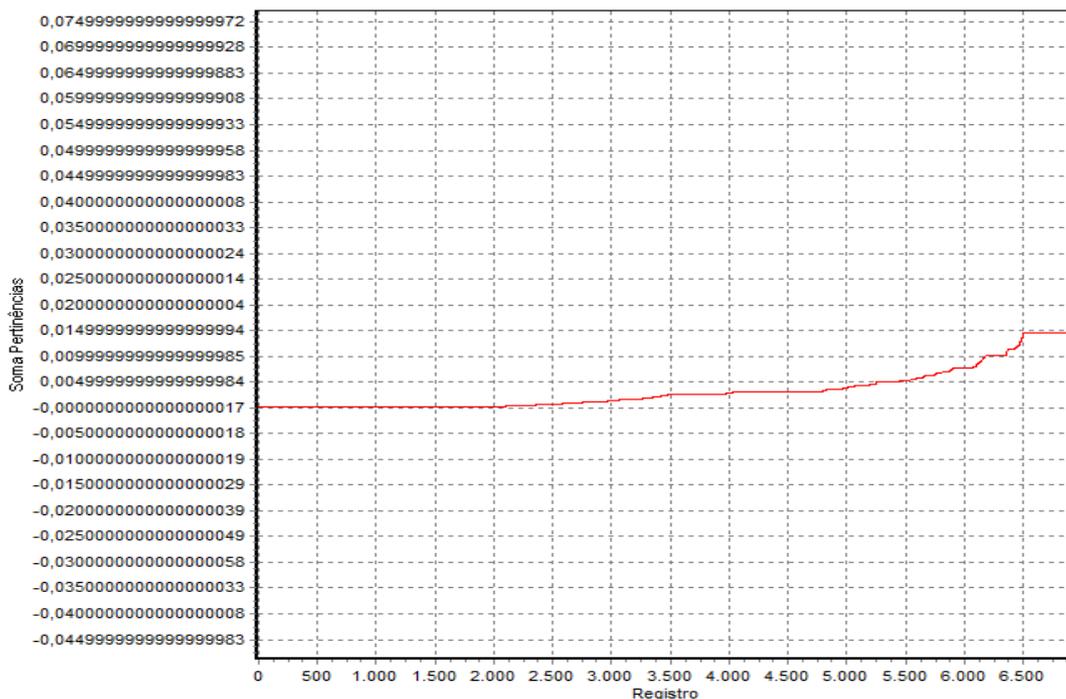


Figura 57 KPCM - Pertinência total (três grupos) - Dados AIH

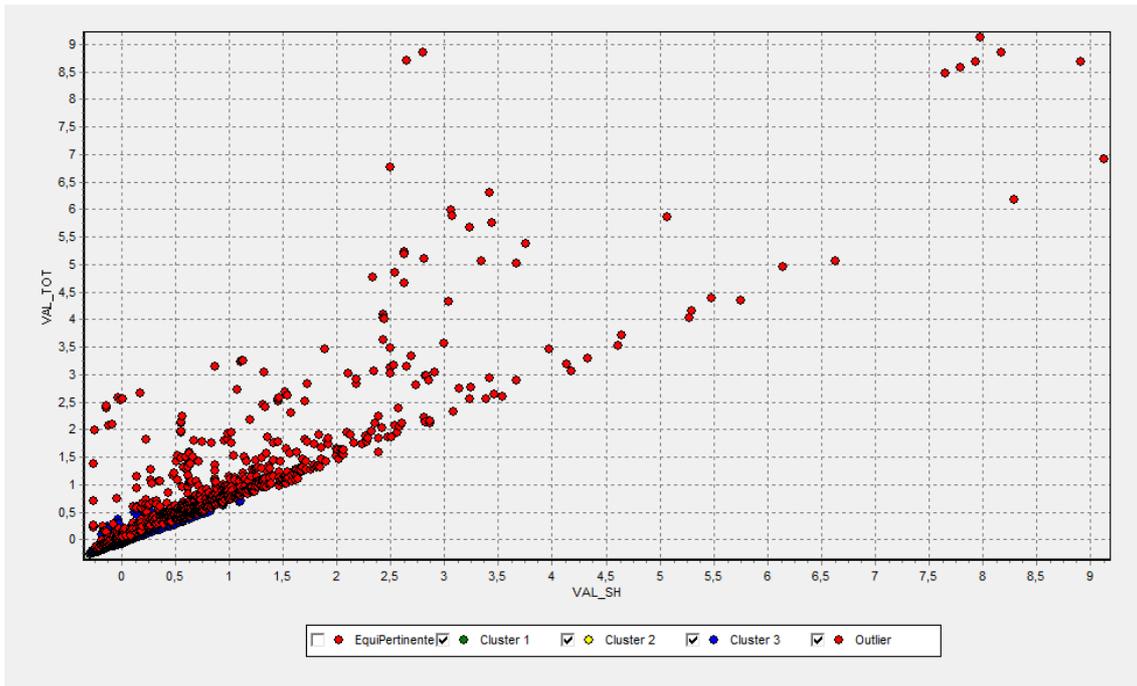


Figura 58 KPCM (três grupos e outliers) - Agrupamento dados AIH

Observa-se um comportamento semelhante para um agrupamento com seis “clusters”. Não há outliers com limite proposto, conforme vê-se na figura 59, mas, aumentando-se recursivamente, da mesma forma como feito anteriormente, este limite, obtém-se 452 dados anômalos ao imputar o valor 0,005, com  $\sigma = 0,707$ , e 1624 registros divergentes com  $\sigma = 1$ . Os dados anômalos seguem representados na figura 60.

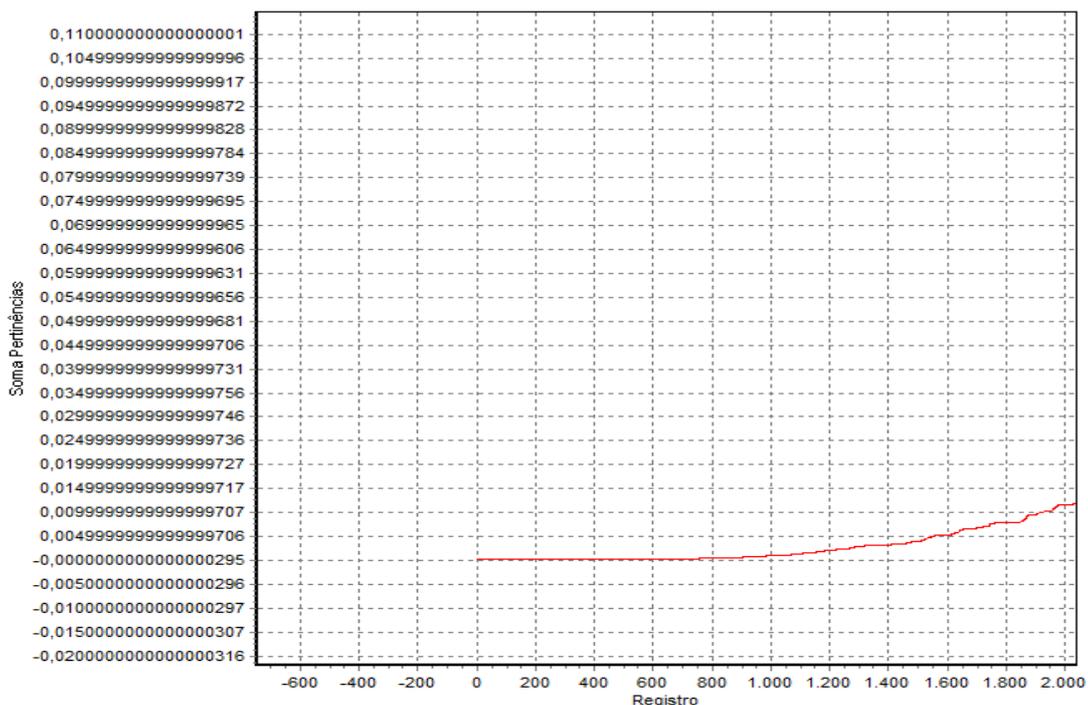
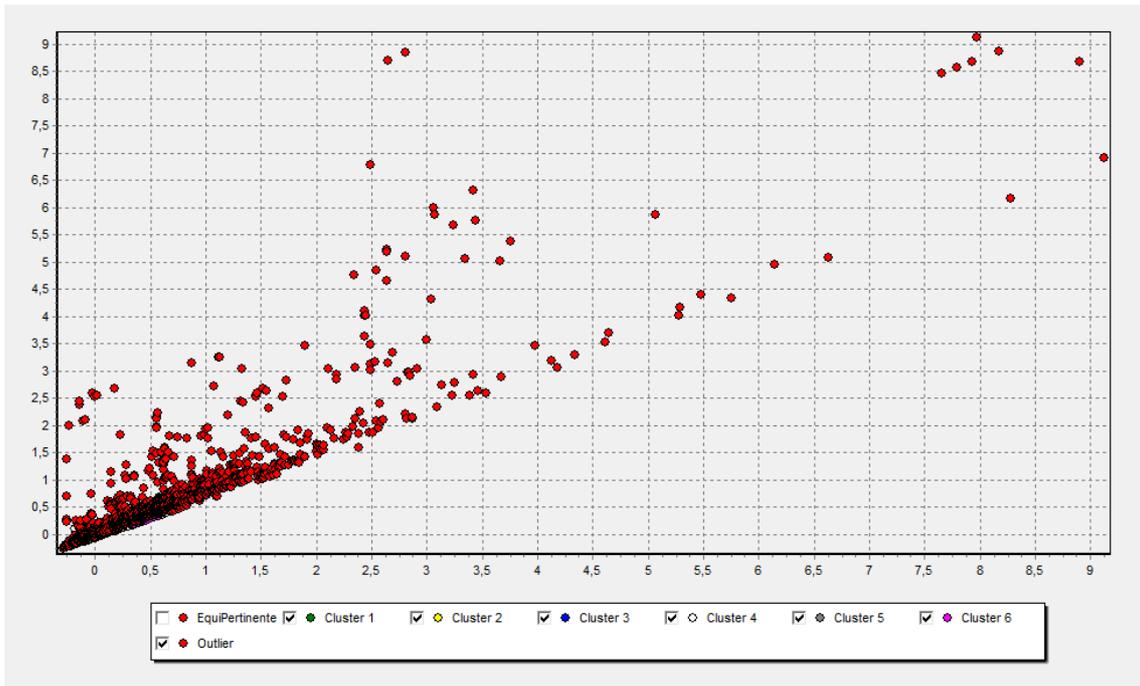
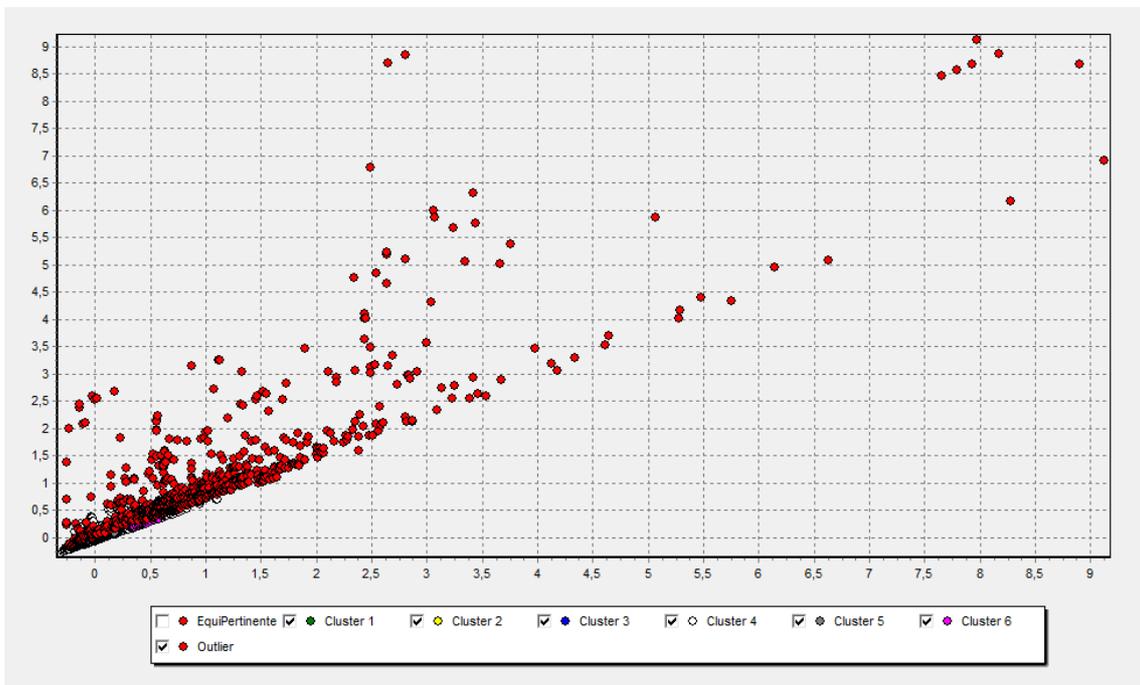


Figura 59 KPCM - Pertinência total (seis grupos) - Dados AIH



**Figura 60** KPCM -(seis grupos e outliers) - Agrupamento Dados AIH

Assim, embora as quantidades de dados anômalos sejam diferentes, graficamente vislumbra-se que há intersecções entre estes resultados, o que pode ser visto ao se comparar com os 889 *outliers* obtidos ao se agrupar em quatro grupos, conforme exposto na figura 61 abaixo:



**Figura 61** KPCM (quatro grupos e outliers) - Agrupamento Dados AIH

## 6.2 Aplicação de Máquina de Vetor Suporte com classe singular (SVM – One Class).

A técnica de máquina de vetor suporte tem demonstrado ser de grande utilidade em várias áreas, produzindo resultados iguais ou superiores a outros classificadores. Tal foi projetado para problemas com duas classes e buscar-se-ia obter um hiperplano que maximizasse a separação entre estas. Entretanto, há casos onde somente se dispõe de um conjunto com uma classe conhecida. Por exemplo, em um problema de reconhecimento do numeral três manuscrito, onde há apenas amostras deste numeral manuscrito, sem outras que demonstrassem o que não é a inscrição deste numeral. Este se transforma em um problema de classe única, que requer separar do restante do conjunto (CHAN, Y. H., 2004).

Esta técnica é categorizada como de classificação de dados, entretanto, busca separar uma classe de objetos de outros rotulados como *outliers*, dispensando o uso de conjunto de dados de treinamento. Tal fato leva-nos a concluir que se trata de um problema de aprendizado não supervisionado e, portanto, mais próximo do conceito de agrupamento.

Apesar de SCHÖLKOPF, B e SMOLA, A. J. (2002) afirmarem que o parâmetro  $c$  não afeta a segregação dos dados, verificou-se um aumento na taxa de falsos alarmes, quando se diminuía o valor deste. Assim, fixar-se-á o parâmetro  $c$  em 0,5, face ter-se obtido baixo percentual de erro com este valor. Assim, fazendo-se variar apenas o parâmetro  $\nu$ , responsável pelo alargamento ou redução do raio da hiperesfera, o que pode evidenciar ou incorporar dados divergentes, chega-se aos resultados da aplicação expostos na tabela 44 a seguir:

**Tabela 44.** Quantidade de dados normais e anômalos (SVM)

$\nu$	N. de registros encontrados	
	Classe 1	Classe -1
0,5	15077	14922
0,25	22755	7244
0,125	24341	5658
0,0625	29233	766
0,03125	25688	520
0,000033	29973	26

Conforme proposto por SCHÖLKOPF, B e SMOLA, A. J. (2002) o parâmetro  $\sigma$  ficou fixado em 0,5 e permitiu-se uma variação de valores de  $\nu$  menores que 0,5. Assim, estes dados ficam separados conforme se expõe nas figuras 62 a 66 abaixo:

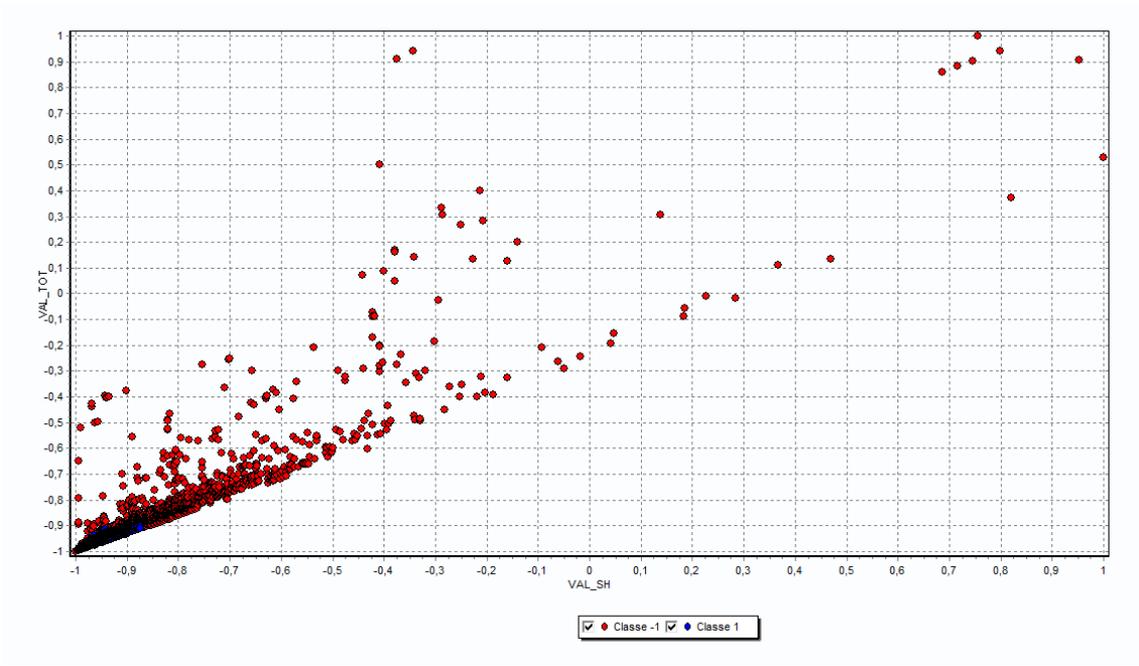


Figura 62 SVM-One Class ( $\nu = 0,5$ ) - Dados AIH

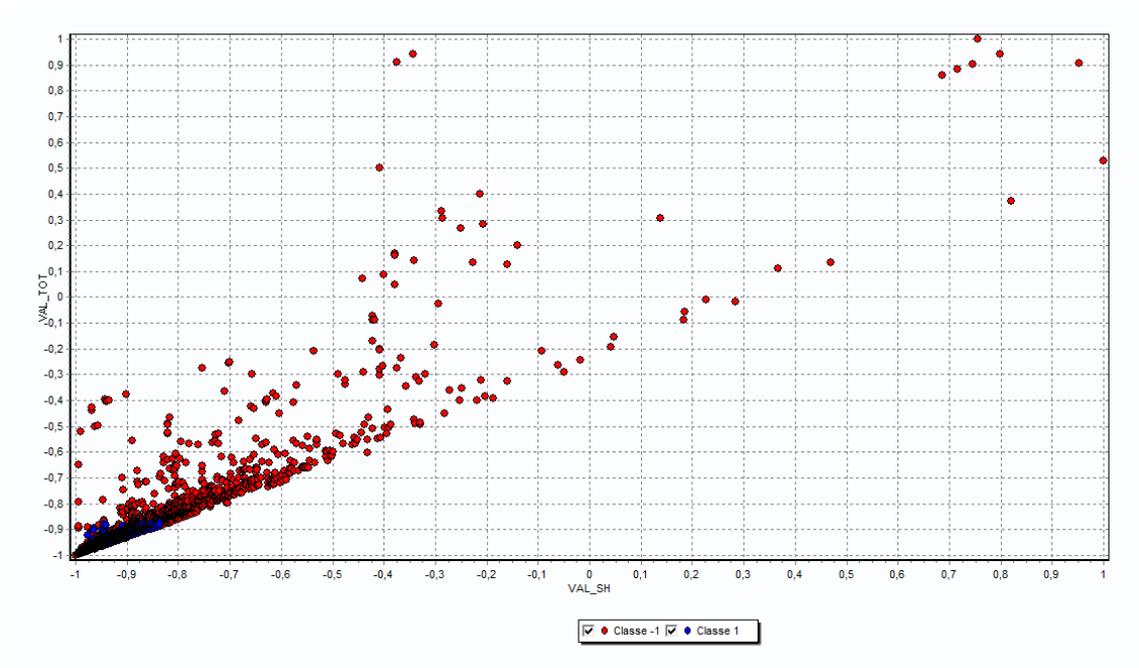


Figura 63 SVM-One Class ( $\nu = 0,25$ ) - Dados AIH

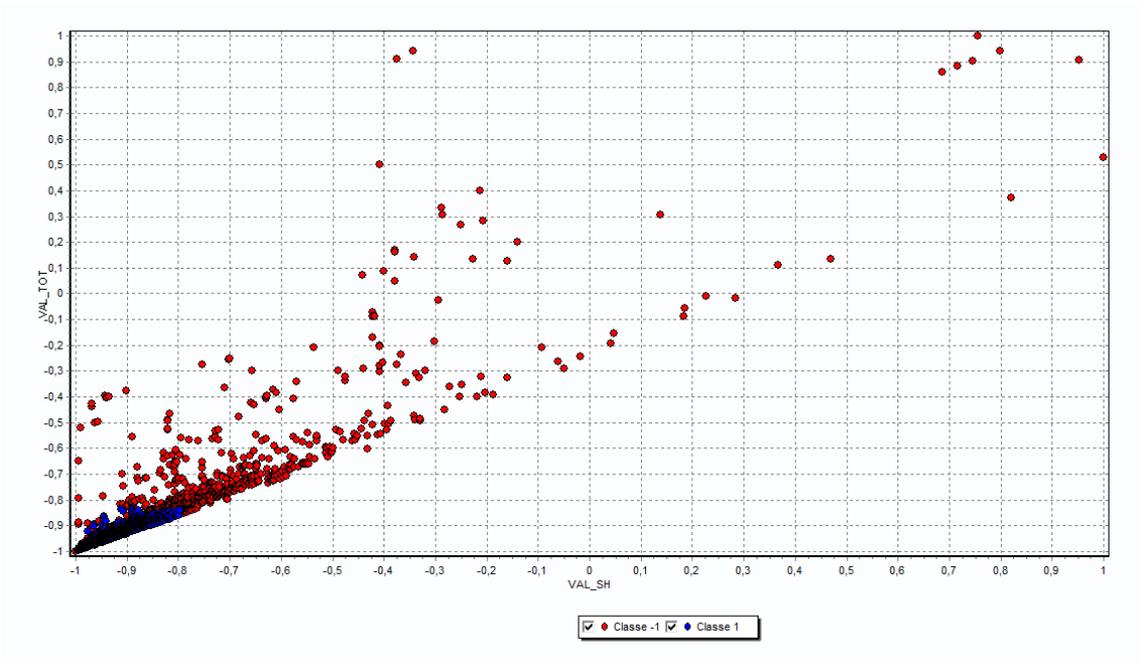


Figura 64 SVM-One Class ( $\nu = 0,125$ ) - Dados AIH

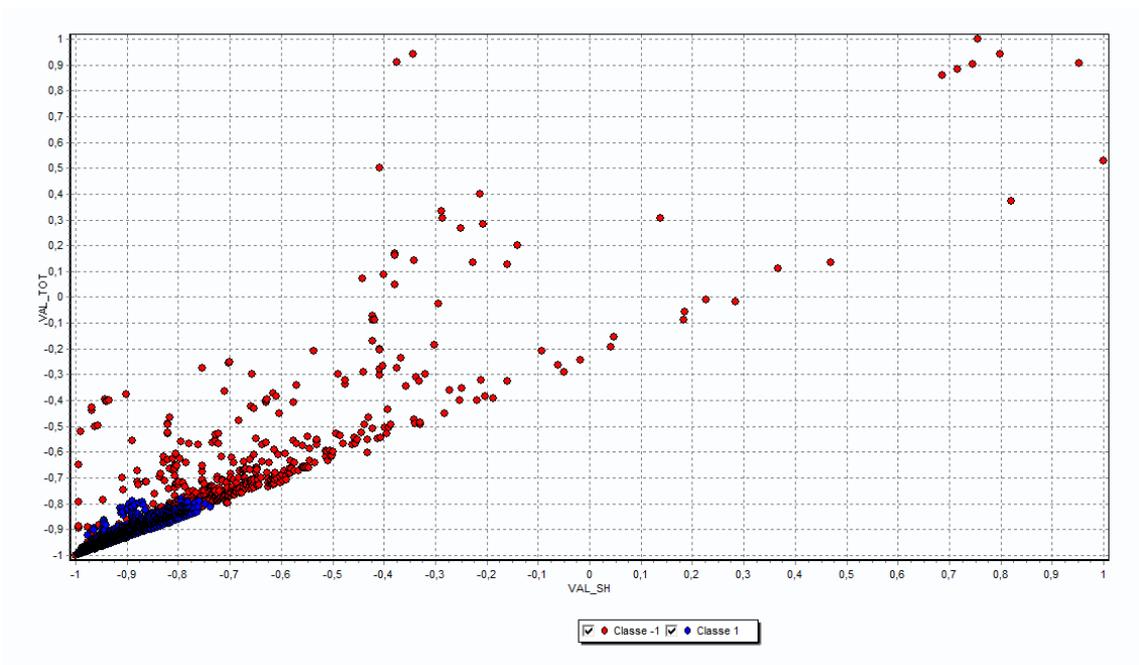


Figura 65 SVM-One Class ( $\nu = 0,0625$ ) - Dados AIH

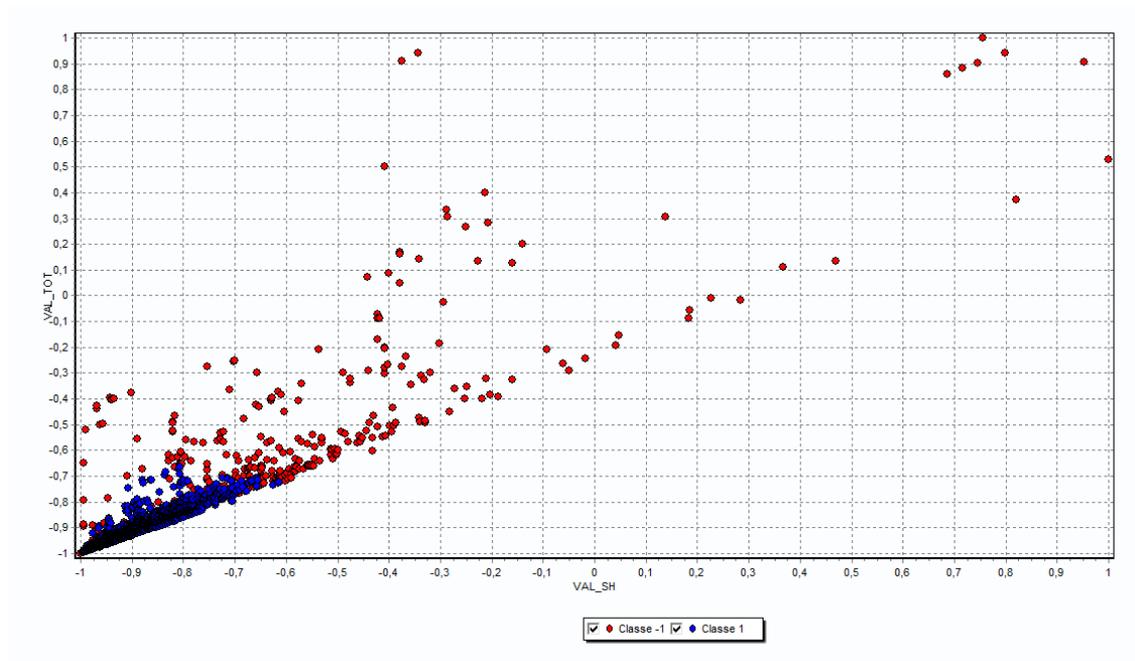


Figura 66 SVM-One Class ( $\nu = 0,03125$ ) - Dados AIH

Interessante notar que os pontos apontados como anômalos com  $\nu$  tendo valor baixo persistem no conjunto de outliers revelados pelo algoritmo quando utilizando-se limites maiores para esse parâmetro.

### 6.3 Aplicativo de Função de Similaridade Média

Este método guarda uma pequena semelhança com o “*K Nearest Neighbour*” (KNN), apesar deste aplicar-se às tarefas supervisionadas de classificação. Neste último, busca-se classificar os dados baseado na distância média entre um ponto e seus K pontos vizinhos mais próximos (VOULGARIS, Z.; MAGOULAS, G, 2008).

Presentes as dificuldades de se determinar o valor de k e do desconhecimento das classes de dados, propõe-se um método que, independente de qualquer conhecimento acerca do conjunto de dados, possa evidenciar dados anômalos baseado na distância média de um ponto a todos os demais, aplicando-se no cálculo uma função Kernel gaussiana.

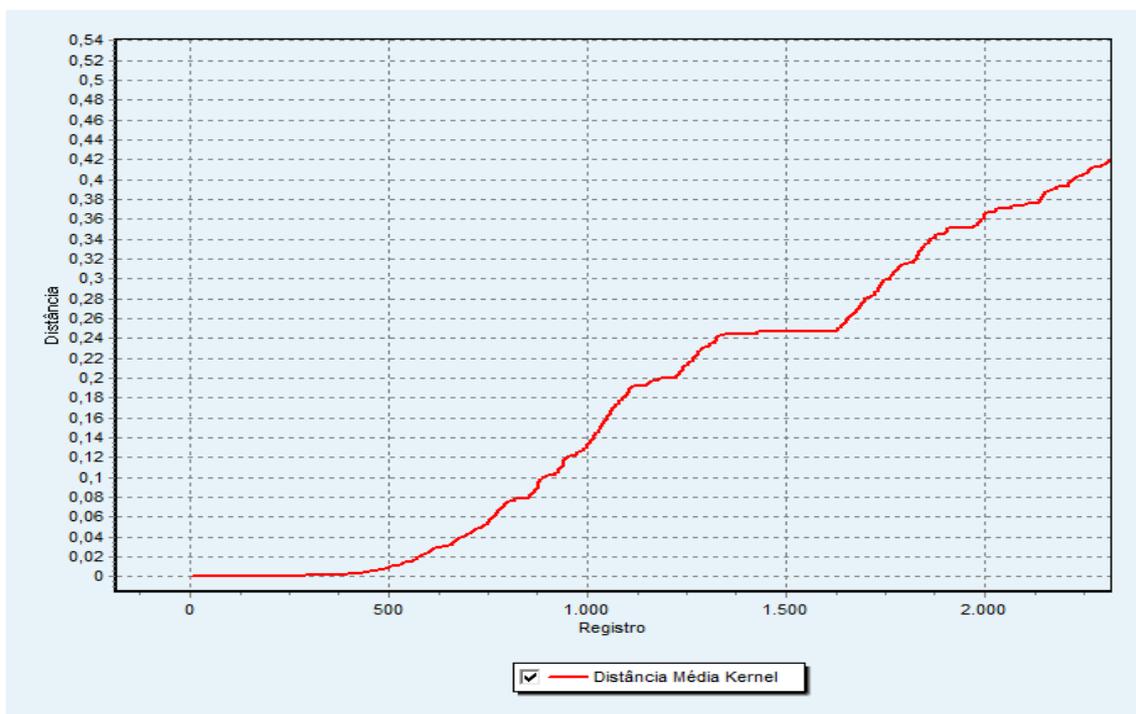
A função Kernel necessita de um parâmetro ( $\sigma$ ) para estabelecer a distância no espaço característico, que é aplicado de forma inversa no aplicativo LIBSVM, cuja compatibilização está exposta na tabela 45.

**Tabela 45.** Correspondência de parâmetros

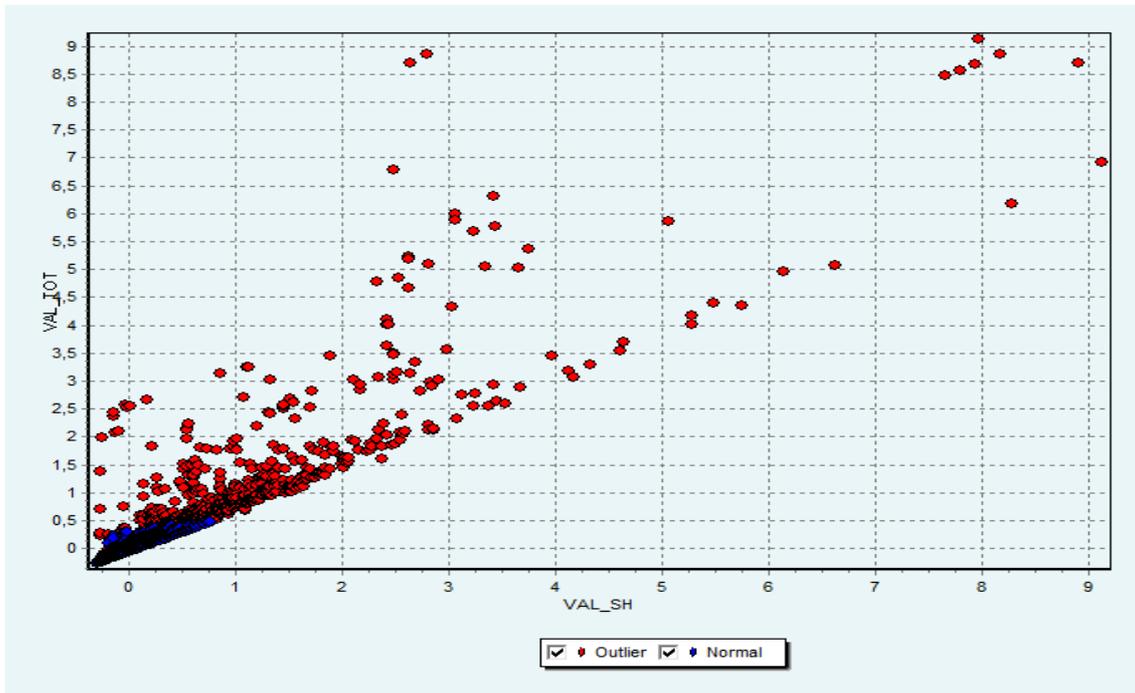
LIBSVM	Função Distância Média
$\gamma$	$\sigma$
2	0,25
1	0,5
0,5	1
0,25	2
0,125	4

Aplicando-se à base de dados AIH, ter-se-á:

- a) Instando o parâmetro  $\sigma = 0,25$ , verifica-se (figura 67) que após uma distância média de 0,24 a curva apresenta um “degrau” e torna a crescer, abaixo do qual contam 1636 registros considerados anômalos (figura 68):

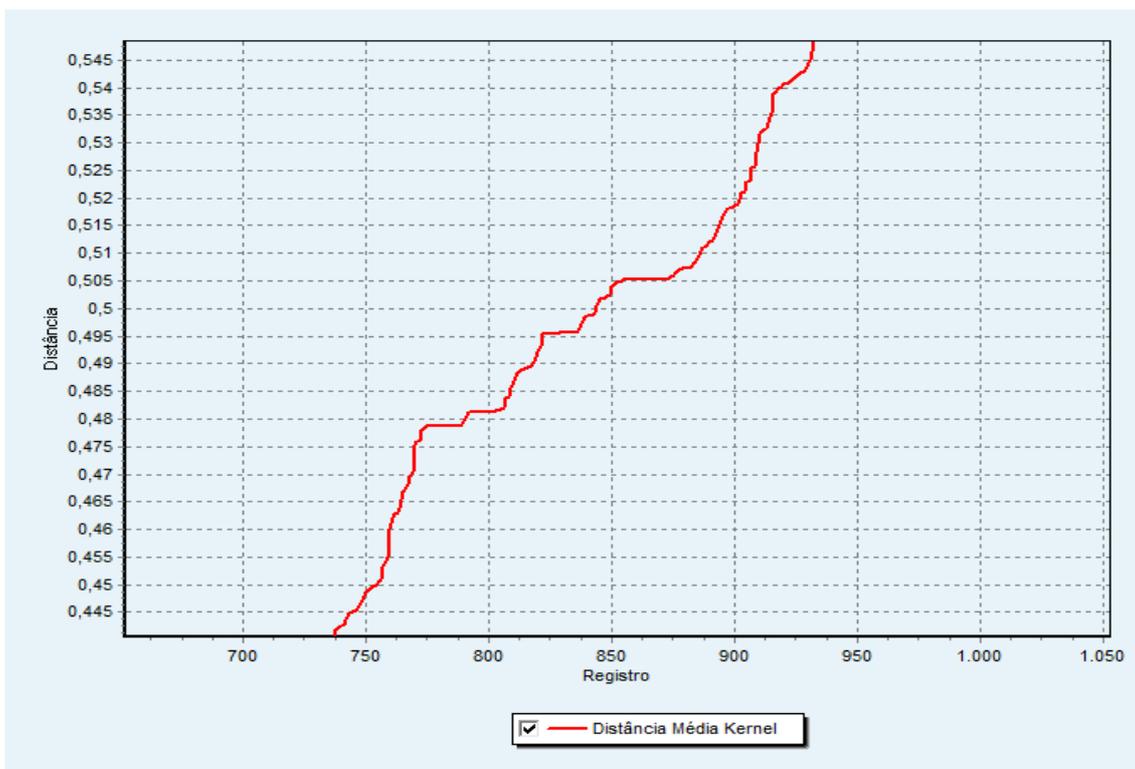


**Figura 67** Função de Similaridade Média ( $\sigma = 0,25$ ) - Dados AIH

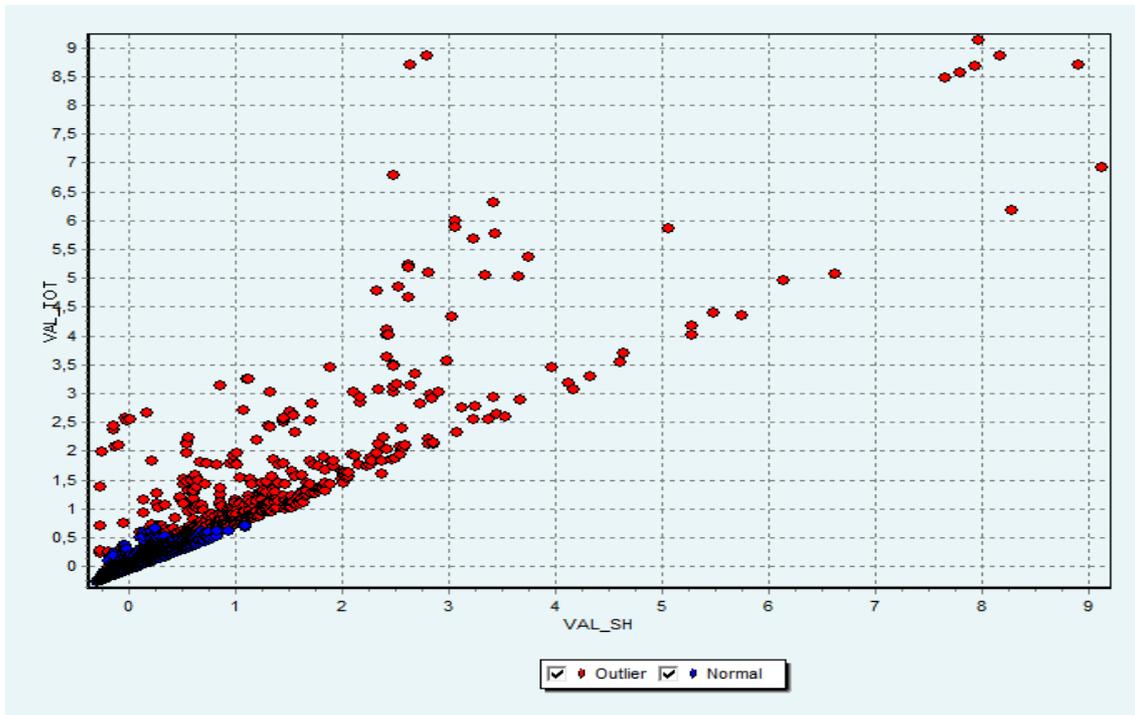


**Figura 68** Separação dos dados - Função de Similaridade Média ( $\sigma = 0,25$ ) - Dados AIH

- b) Instando o parâmetro  $\sigma = 0,5$ , vê-se que, no limite da distância média de 0,48 (observe o primeiro “degrau” que a curva apresenta na figura 69), obteve-se 797 outliers, evidenciados na figura 70.

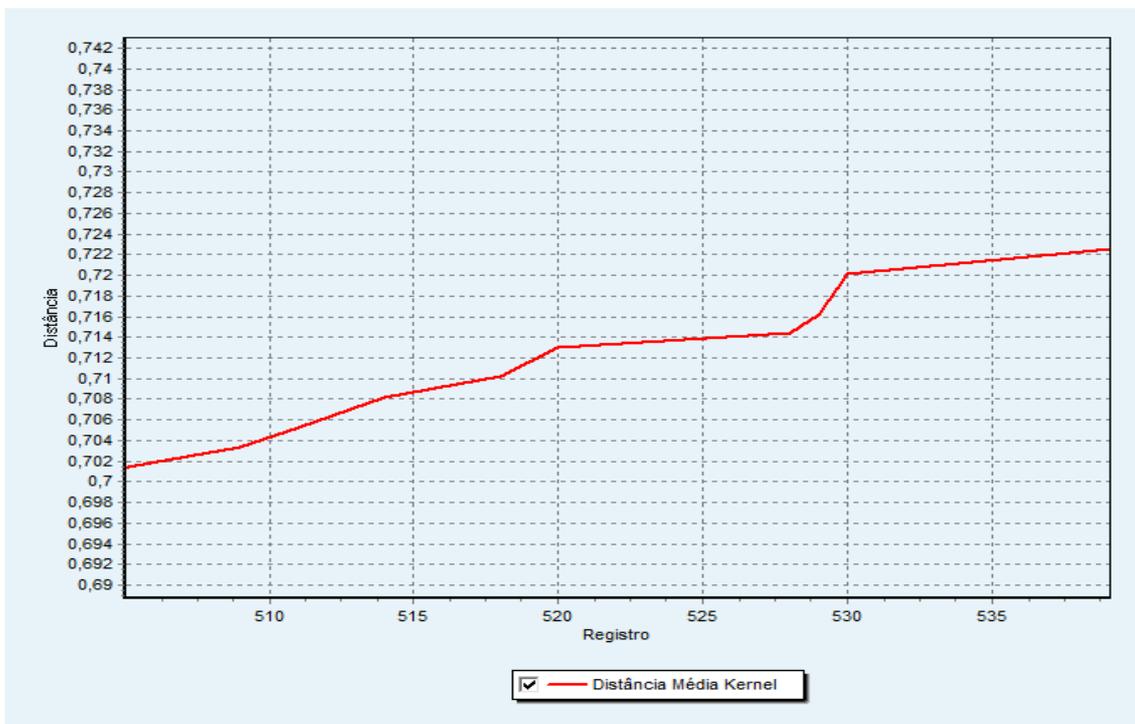


**Figura 69.** Função de Similaridade Média ( $\sigma = 0,5$ ) - Dados AIH

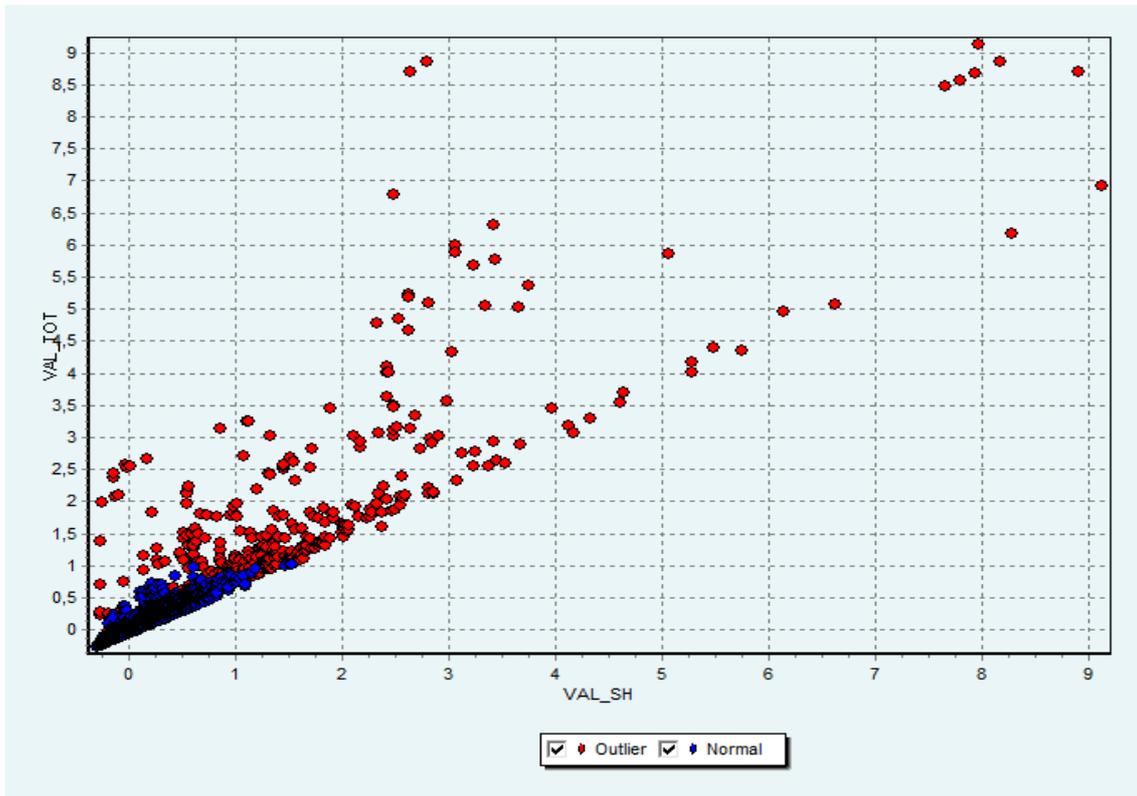


**Figura 70** Separação dos dados - Função de Similaridade Média ( $\sigma = 0,5$ ) - Dados AIH

- c) Instando o parâmetro  $\sigma = 1$  observa-se que no limite da distância média 0,71, evidenciam-se 519 dados anômalos (veja o “degrau” que a curva apresenta na figura 71). O outliers são apresentados no gráfico da figura 72 :

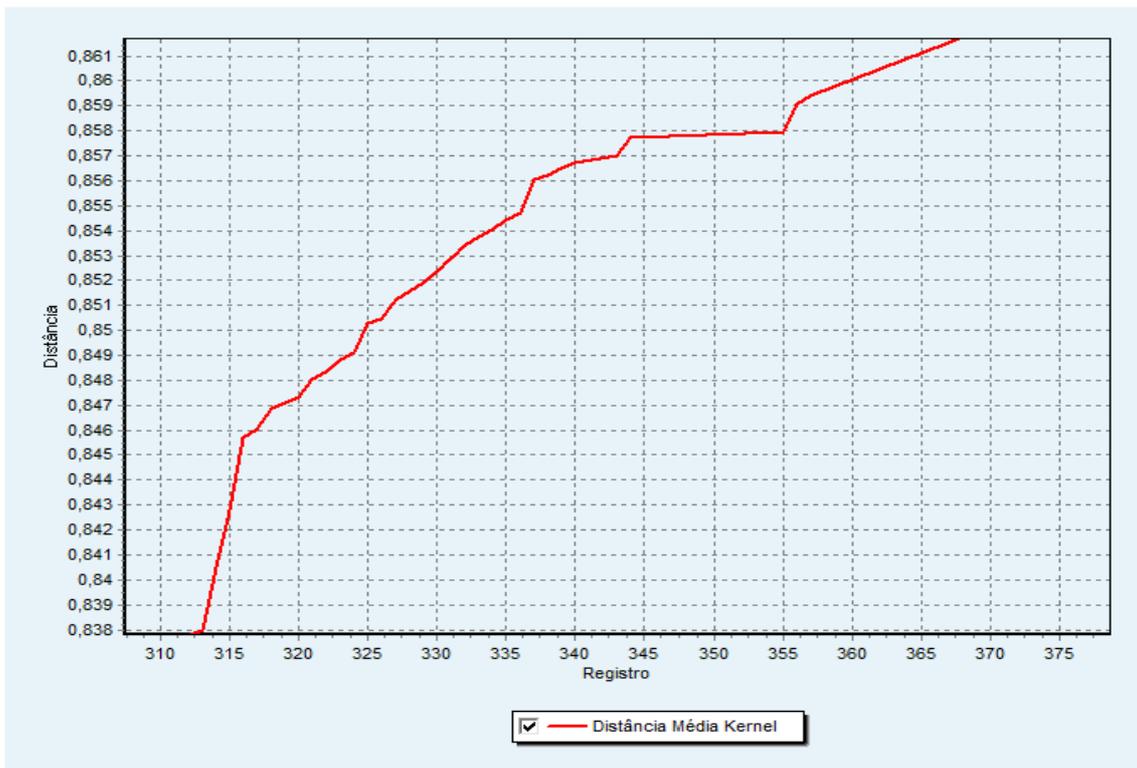


**Figura 71** Função de Similaridade Média ( $\sigma = 1$ ) - Dados AIH



**Figura 72** Separação dos dados - Função de Similaridade Média ( $\sigma = 1$ ) - Dados AIH

- d) Instando o parâmetro  $\sigma = 2$ , observa-se que no limite da distância de 0,858, chega-se a 324 dados anômalos, como evidenciam as figuras 73 e 74:



**Figura 73** Função de Similaridade Média ( $\sigma = 2$ ) - Dados AIH

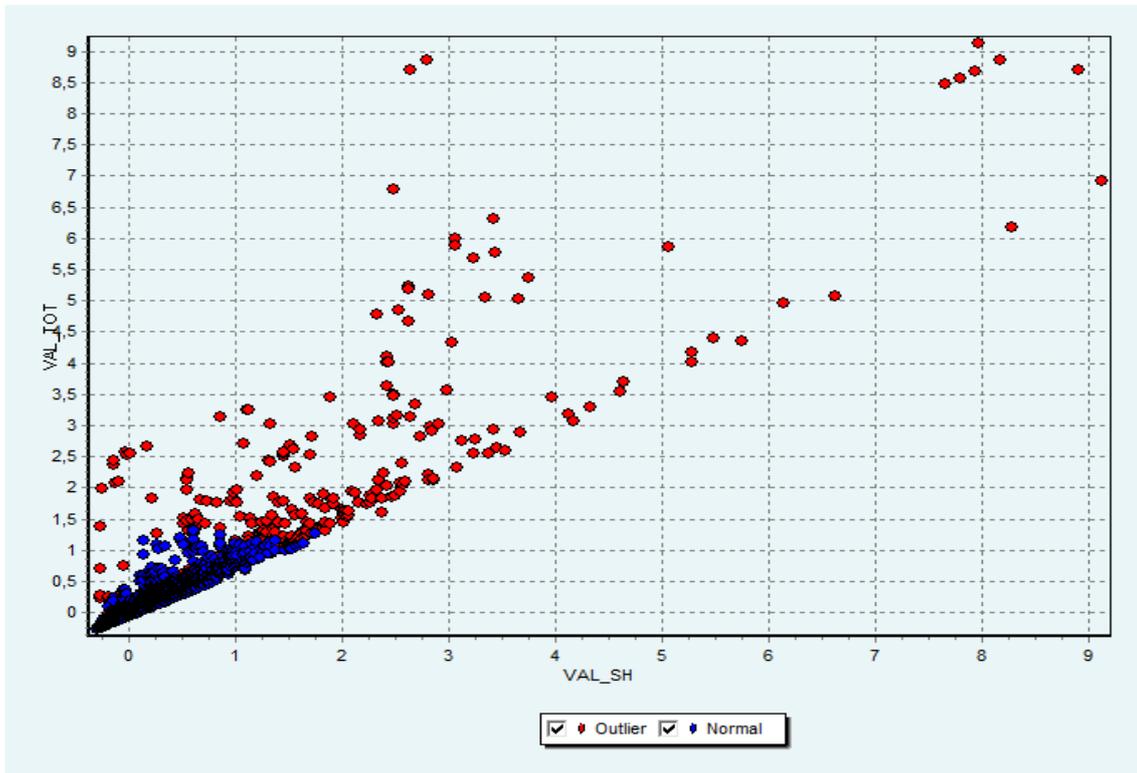


Figura 74 Separação - Função de Distância Média ( $\sigma = 2$ ) - Dados AIH

- e) Instando o parâmetro  $\sigma = 4$ , o limite da distância de 0,945 determina a existência de 90 dados anômalos, conforme vê-se nos gráficos das figuras 75 e 76:

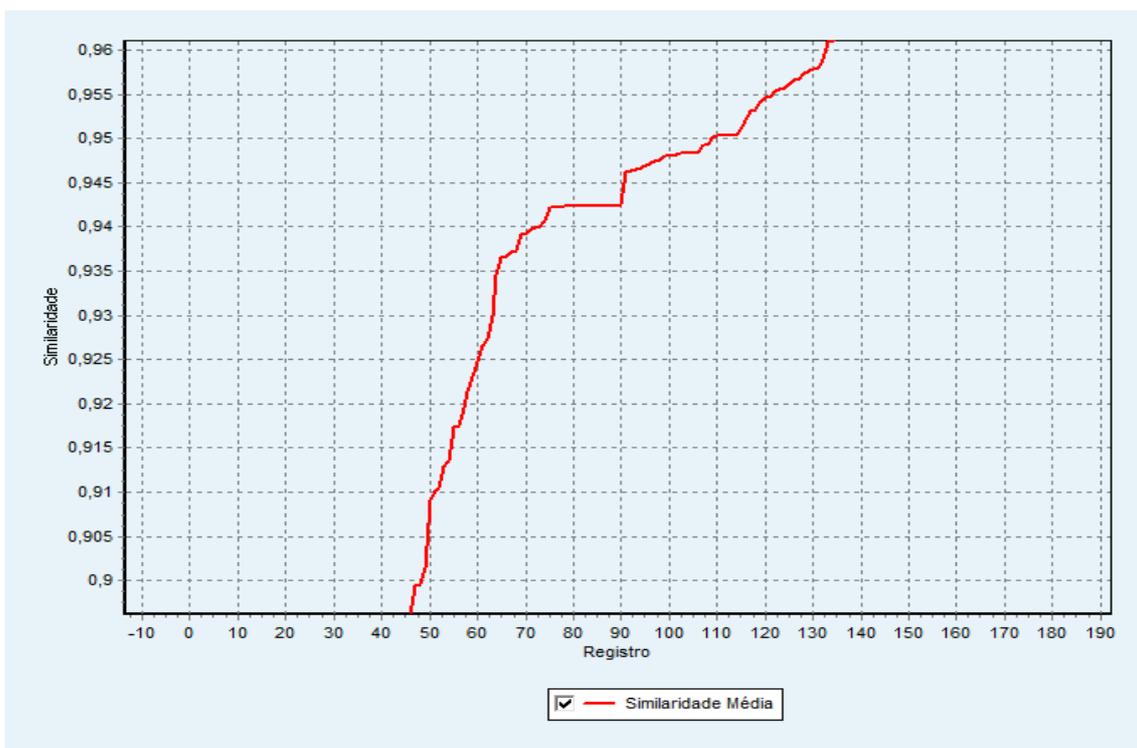


Figura 75 Função de Similaridade Média ( $\sigma = 4$ ) - Dados AIH

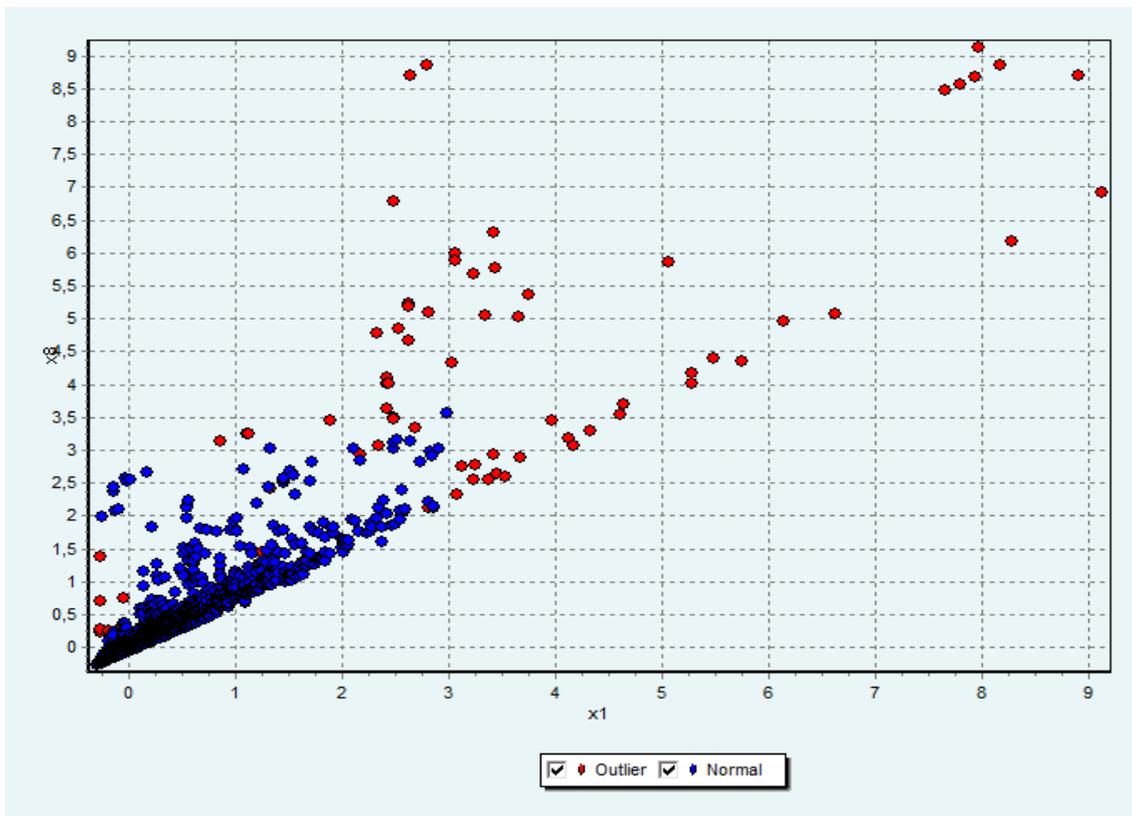


Figura 76 Separação - Função de Distância Média ( $\sigma=4$ ) - Dados AIH

Os dados de AIH submetidos, de acordo com o valor de  $\sigma$ , ficam estratificados segundo o quadro da tabela 46 abaixo:

Tabela 46. Quantidade de dados normais e anômalos (FSM)

$\sigma$	Quantidade de Dados Normais	Quantidade de dados anômalos
0,25	28363	1636
0,5	29202	797
1	29480	519
2	29675	324
4	29009	90

## 6.4 DISCUSSÃO

Não há, neste conjunto de dados, registros sabidamente anômalos, o que impede estabelecer o grau de eficácia dos métodos ao evidenciar *outliers*. Entretanto, os testes de precisão efetuados anteriormente permitem afirmar que têm alta capacidade de separar os dados anômalos.

Face às quantidades diferentes, é interessante investigar o grau de concordância intra-método, confrontando-se os resultados diferentes do mesmo método, e inter-método, comparando-se os resultados obtidos pelos três métodos.

Verificando-se as coincidências de registros rotulados como anômalos pelo método de agrupamento nebuloso, considerando quatro *clusters*, variando o valor do parâmetro  $\sigma$  da função Kernel, observa-se a ocorrência das quantidades de dados anômalos exposta na tabela 47 abaixo:

**Tabela 47.** Quantidade de dados anômalos (KPCM)

$\sigma$	0,25	0,5	1	2	4
<b>Quantidade de dados anômalos</b>	889	1739	1764	1717	1695

Confrontando os registros apontados, segundo cada valor de  $\sigma$ , verifica-se a concordância no assinalamento do registro como *outlier* conforme expõe a tabela 48, isto é, os 889 registros apontados como anômalos, ao se atribuir o valor 0,707 ao parâmetro  $\sigma$ , estão presentes nos conjuntos de dados anômalos determinados pela atribuição dos outros valores a esse parâmetro. Dos 1739 registros anômalos apontados com  $\sigma = 1,000$ , 1737 encontram-se presentes no conjunto encontrado quando  $\sigma$  assume o valor 1,414. Ou seja, o conjunto com maior número de elementos contém os elementos dos conjuntos menores.

**Tabela 48.** Concordância intra-método (KPCM)

$\sigma$	0,5	1	2	4
<b>0,25</b>	889	889	889	889
<b>0,5</b>		1737	1712	1693
<b>1</b>			1717	1694
<b>2</b>				1693

Há uma quase unanimidade na classificação, sendo que a concordância pode ser expressa nos seguintes percentuais expostos na tabela 49:

**Tabela 49.** Percentual de concordância intra-método (KPCM)

$\sigma$	0,5	1	2	4
<b>0,25</b>	100,00%	100,00%	100,00%	100,00%
<b>0,5</b>		99,88%	99,71%	99,88%
<b>1</b>			100,00%	99,94%
<b>2</b>				99,88%

Observando-se o resultado apresentado pelo algoritmo *SVM - One Class*, obtém-se a seguinte síntese exposta na tabela 50:

**Tabela 50.** Quantidade de dados anômalos (*SVM – One Class*)

$\nu$	Qt. Dados Anômalos
0,5	14.922
0,25	7.244
0,125	5.658
0,0625	766
0,03125	520

Verifica-se que os *outliers* estabelecidos pelo algoritmo, ao utilizar  $\nu = 0,03125$  (menor número), estão presentes no conjunto obtido quando se atribui o valor 0,0625 ao parâmetro  $\nu$ , e assim sucessivamente. Assim a quantidade de registros existentes na intersecções dos conjuntos fica sintetizada na tabela 51:

**Tabela 51.** Quantidade de dados anômalos coincidentes (*SVM – One Class*)

$\nu$	0,25	0,125	0,0625	0,03125
0,5	7.244	5.658	766	520
0,25		5.658	766	520
0,125			766	520
0,0625				520

A concordância entre as classificações providas pelo aplicativo, de acordo com cada valor do parâmetro  $\nu$ , é sintetizada na tabela 52 abaixo :

**Tabela 52.** Percentual de concordância intra-método (*SVM – One Class*)

$\nu$	0,25	0,125	0,0625	0,03125
0,5	100,00%	100,00%	100,00%	100,00%
0,25		100,00%	100,00%	100,00%
0,125			100,00%	100,00%
0,0625				100,00%

O aplicativo de função de Similaridade Média determinou as seguintes quantidades de dados anômalos expressas na tabela 53 de acordo com o valor do parâmetro  $\sigma$  da função kernel:

**Tabela 53.**Quantidade de dados anômalos (FSM)

$\sigma$	Valor Limite de distância	Quantidade de dados anômalos
<b>0,25</b>	0,2525	1636
<b>0,5</b>	0,479	790
<b>1</b>	0,712	519
<b>2</b>	0,85	324
<b>4</b>	0,945	90

Executando o mesmo procedimento anterior para os resultados providos pelo aplicativo, com a variação do parâmetro da função kernel, chega-se aos valores de concordância intra-método exibidos na tabela 54:

**Tabela 54.**Concordância intra-método (FSM)

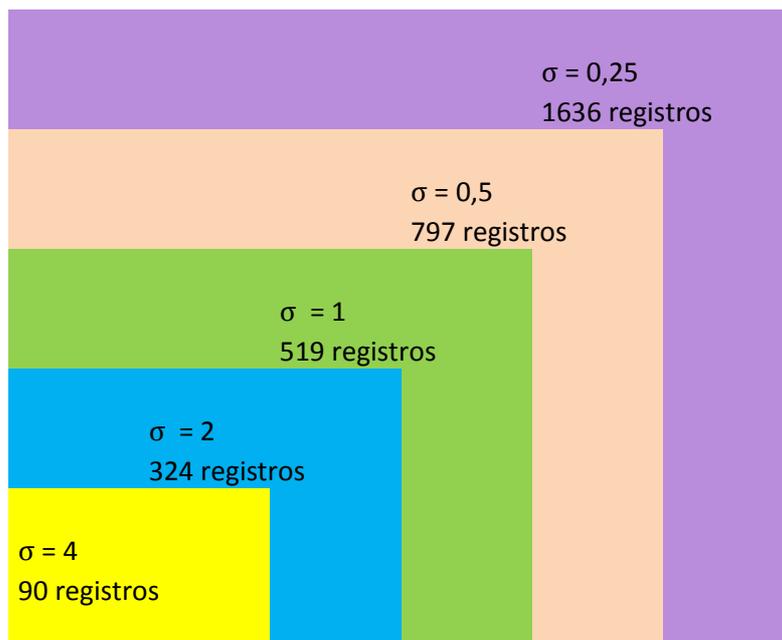
$\sigma$	<b>0,5</b>	<b>1</b>	<b>2</b>	<b>4</b>
<b>0,25</b>	790	519	324	90
<b>0,5</b>		519	324	90
<b>1</b>			324	90
<b>2</b>				90

Vê-se que há uma concordância total nos registros evidenciados pelo método, cujos percentuais de concordância ficam expressos na tabela 55 abaixo:

**Tabela 55.**Percentual de concordância intra-método (FSM)

$\sigma$	<b>0,5</b>	<b>1</b>	<b>2</b>	<b>4</b>
<b>0,25</b>	100,00%	100,00%	100,00%	100,00%
<b>0,5</b>		100,00%	100,00%	100,00%
<b>1</b>			100,00%	100,00%
<b>2</b>				100,00%

Desses resultados, vê-se que a concordância intra-método de todos algoritmos é praticamente total, onde os registros anômalos em menor quantidade estão presentes no conjunto com quantidade imediatamente superior e assim sucessivamente, como indicado na figura 77 abaixo:



**Figura 77.** Interseção dos Registros anômalos (FSM)

Há que se testar também a concordância inter-métodos, isto é, verificar se os registros indicados como anormais por um método também o são pelos outros. Por outras palavras, verificar se os 90 registros classificados pelo método de Função de Similaridade Média (vide figura 77 acima) estão presentes no conjunto de dados anômalos evidenciados pelos algoritmos de Agrupamento Nebuloso e de Máquina de Vetor Suporte.

Assim, ao se confrontar os resultados dos aplicativos de Função de distâncias médias e agrupamentos nebulosos, a tabela 56 expõe as quantidades de registros com classificação igual:

**Tabela 56.** Quantidade de registros com classificação igual (FSM x KPCM)

			Agrupamento Nebuloso (KPCM)				
			$\sigma$				
			0,25	0,5	1	2	4
	$\sigma$	N. de outliers	889	1739	1764	1717	1695
Função De Similaridade Média (FSM)	0,25	1636	889	1625	1627	1626	1626
	0,5	797	790	790	790	790	790
	1	519	519	519	519	519	519
	2	324	324	324	324	324	324
	4	90	90	90	90	90	90

Vê-se que os 889 outliers estabelecidos pelo aplicativo KPCM, com  $\sigma = 0,707$ , estão presentes no conjunto de 1.636 outliers determinado pelo algoritmo FSM, com mesmo

valor do parâmetro. Os 790 dados anômalos determinados pelo aplicativo FSM, com  $\sigma = 1,0$ , estão presentes no conjunto de 889 outliers evidenciados pelo algoritmo KPCM com  $\sigma = 0,707$  e assim sucessivamente.

Esses valores ficam mais claramente entendidos quando se observa o índice de concordância, conforme exposto na tabela 57 abaixo:

**Tabela 57.** Percentual de registros com classificação igual (FSM x KPCM)

		Agrupamento Nebuloso (KPCM)				
		$\sigma$	0,25	0,5	1	2
Função De Similaridade Média (FSM)	0,25	100,00%	99,33%	99,45%	99,39%	99,39%
	0,5	99,12%	99,12%	99,12%	99,12%	99,12%
	1	100,00%	100,00%	100,00%	100,00%	100,00%
	2	100,00%	100,00%	100,00%	100,00%	100,00%
	4	100,00%	100,00%	100,00%	100,00%	100,00%

Observa-se que praticamente todos os registros apontados como anômalos pelo método de Função de Similaridade Média são classificados de igual forma pelo algoritmo de Agrupamento Nebuloso.

No confronto dos resultados entre os aplicativos de máquina de vetor suporte, classe única (SVM One-Class) e de Função de Similaridade Média, chega-se às quantidades de registros classificados de mesma maneira na tabela 58 abaixo:

**Tabela 58.** Quantidade de registros com classificação igual (FSM x SVM – One Class)

			SVM One-Class				
			$\nu$				
			0,5	0,25	0,125	0,0625	0,03125
$\sigma$	N. de outliers	14.922	7.244	5.658	766	520	
Função De Similaridade Média (FSM)	0,25	1636	1636	1612	1125	637	394
	0,5	797	790	790	783	620	384
	1	519	519	519	519	510	366
	2	324	324	324	324	324	316
	4	90	90	90	90	90	90

Transformando esse resultado em valores relativos, obtém-se os seguintes índices de coincidência expostos na tabela 59:

**Tabela 59.** Percentual de registros com classificação igual (FSM x SVM – One Class)

	$\sigma$	SVM One-class				
		$\nu$				
		0,5	0,25	0,125	0,0625	0,03125
Função De Similaridade Média	0,25	100,00%	98,53%	68,77%	83,16%	24,14%
	0,5	99,12%	99,12%	98,24%	80,94%	48,31%
	1	100,00%	100,00%	100,00%	98,27%	70,71%
	2	100,00%	100,00%	100,00%	100,00%	97,84%
	4	100,00%	100,00%	100,00%	100,00%	100,00%

Para o valor  $\nu = 0,5$  vê-se uma concordância quase total, mas deve-se ressaltar que o aplicativo SVM considera quase a metade dos dados como anômalos e, dessa forma, a desproporção entre esses e o número de dados anormais exposto pela Função de Similaridade Média leva a uma concordância alta. É interessante notar que, ao se comparar conjuntos com números de elementos semelhantes, obtém-se praticamente uma concordância adequada quantos aos registros anômalos, como descreve a tabela 60 abaixo:

**Tabela 60.** Concordância na separação de dados (FSM x SVM One-Class)

	$\sigma$	N. de dados Anômalos	SVM One-Class ( $\nu = 0,0625$ )
			766
Função De Similaridade Média (FSM)	1,000	797	80,94%
	1,414	519	98,27%

Conclui-se, portanto, que os métodos analisados são eficazes na evidenciação do dado anômalo.

O algoritmo de Máquina de Vetor Suporte é um método sobejamente aplicado às tarefas de classificação de dados e, mormente no caso do SVM - One Class, constitui um método com aplicativo amplamente testado, capaz de separar um conjunto em dois grupos, sendo um de especial interesse neste estudo, a classe de dados anômalos.

O agrupamento nebuloso *Possibilistic C-means* (KPCM), provido com função kernel, é reconhecido como uma forma de agrupar eficazmente o conjunto de dados em diversos grupos, com a característica particular de que a soma das pertinências do registro próxima de zero o qualifica como uma dado anômalo. A busca por este dado

facilita o uso do algoritmo, pois a determinação prévia do número de grupos fica flexibilizada.

Assim, posto dois métodos consagrados, incorpora-se mais uma ferramenta na busca pelo dado divergente, a Função de Similaridade Média (FSM), de fácil construção e aplicação, onde, por meio de uma função kernel guassiano, verifica-se que os registros com valores pequenos são prováveis *outliers*.

Esta nova ferramenta confrontada com os dois primeiros métodos, demonstra ser um método robusto, pois há uma concordância adequada dos seus resultados obtidos com aqueles, quando da submissão dos mesmos dados.

Neste algoritmo, pode-se ordenar crescentemente os registros segundo o valor de sua similaridade, tendo-se, desta forma, uma gradação da dissimilaridade dos registros em relação ao conjunto. Assim, a tarefa do auditor fica simplificada, posto que ficam evidenciados quais registros devem ser verificados. A decisão de quantos destes dados serão analisados fica, portanto, subordinada à capacidade de trabalho do auditor ou da equipe e ao tempo disponível para a tarefa.

O algoritmo KPCM apresenta uma semelhança com FSM no que tange a capacidade de ordenar os registros segundo a soma dos graus de pertinência aos *clusters*, entretanto necessita de dois parâmetros, o valor de  $\sigma$  para a função Kernel e o número de agrupamentos pretendidos, o que torna mais difícil sua aplicação. Esta dificuldade é aumentada, pois tem-se de informar um protótipo com os centros dos *clusters* (de acordo como número pretendido) para dar início à execução do procedimento.

O algoritmo SVM *One-class* resulta na separação dos dados em duas classes, não apresentando gradação entre os dados classificados como *outliers*. O auditor, não conseguindo priorizar os registros segundo sua dissimilaridade, tem de analisar todo o conjunto, o que pode ser inviável face o tempo e pessoal disponíveis. A extração de parcelas dos *outliers* pode levar a dispensar-se registros importantes, posto que decisão é subjetiva e conduziria ao retorno à submissão da experiência dos auditores. Apresenta uma necessidade maior quanto ao número de parâmetros necessários, o valor de  $\sigma$  para a função Kernel e o valor de  $\nu$ , que regula a expansão ou retração da margem de separação entre os dados, cujo balanço não é tarefa fácil.

Assim, pode-se afirmar que o algoritmo FSM é eficaz no reconhecimento de *outliers*. É mais simples na sua execução em virtude de ser especializado no levantamento dos dados anômalos, requerendo apenas um parâmetro de entrada (o

valor de  $\sigma$  para a função Kernel). O auditor é dispensado de conhecer o comportamento dos dados e os conceitos ali implementados, sendo capaz de utilizar tal ferramenta de imediato.

## **7. Conclusão e sugestão para futuros trabalhos.**

A existência de recursos escassos frente a necessidades crescentes é um dilema que envolve todo sistema econômico. O enfrentamento desta tarefa determina maximizar a eficiência no uso desses recursos por meio do estabelecimento de prioridades e objetivos.

A verificação da eficiência envolve uma auditoria operacional, onde são verificados todos os procedimentos executados por um gestor, visando, conforme BITTENCOURT, F.M.R. (2006), saber se a entidade adquire, protege e usa seus recursos sem desperdícios e as causas de eventuais práticas errôneas, além de observar o cumprimento das normas legais.

Segundo KIRKOS, E. et al. (2007) a auditoria é uma tarefa que tem demanda crescente face ao volume financeiro cada vez maior sendo transacionado, bem como a exigência de maior transparência e informação por atores de diferentes áreas. Neste caso, o interesse do auditor é evidenciar dados divergentes dos normais, o que é dificultado, quando se usa procedimentos padrões.

A busca pela eficiência é pesquisada também em áreas públicas, como na gerência da saúde pública, onde se vê tal preocupação evidenciada pelo nascimento da lei de responsabilidade fiscal, onde mais que responsabilizar pessoas, busca diminuir riscos e corrigir desvios que afetem o equilíbrio orçamentário.

Uma dessas áreas é a saúde pública, onde se encontram ações no sentido de buscar a otimização no uso de recursos financeiros, como a Norma de Operação Básica do SUS, que determina a realização de auditorias analítica e operacional a fim de analisar os resultados e propor medidas corretivas no processo de decisão da alocação dos recursos, e o manual de auditoria do SUS, cujo objetivo é buscar meios de racionalização de gastos, evitando e detectando fraudes e malversação de recursos públicos.

Em concordância, VASCONCELOS, M. M. et al. (2002) reconhecem a crônica escassez de recursos e a necessidade de prover suporte aos diversos agentes dessa área no intuito de alavancar a aplicação dos recursos. Nesse mister, ciência & tecnologia apresentam-se como parte imprescindível ao suporte, citando-se o uso de métodos estatísticos, modelos matemáticos e mineração de dados.

Vê-se que esta preocupação está presente em diversos países, como expõem PHUA, C. et al. (2004) e PENG, Y. et al. (2006) na pesquisa às instituições

securitárias norte-americanas; HAWKINS, S. et al. (2001), quando evidenciam os problemas enfrentados pelo sistema público de saúde australiano; ORTEGA, P. A. et al. (2006), na análise dos problemas do sistema de saúde pública do Chile; e LIOU, F. et al. (2008), expondo o sistema de saúde de Taiwan. Todos esses autores evidenciam a necessidade do aumento da eficiência na aplicação de recursos e concluem que a mineração de dados é uma ferramenta de melhor utilidade nesse caso.

Todos estes estudos buscam encontrar uma forma para evidenciar um dado anômalo, que é um registro que desvia marcadamente dos demais e a forma mais popular para sua detecção é o cálculo de distância entre objetos conforme expõem HODGE, J. V. e AUSTIN, J. (2004) e PETROVSKIY, M. I. (2003).

Assim, propôs-se, coerente com esses estudos, utilizar as técnicas de mineração de dados já consagradas, Agrupamento Nebuloso de dados e Máquina de Vetor Suporte – Classe Única (*SVM – One Class*), e a Função de Similaridade Média, uma inovação semelhante ao algoritmo *K nearest neighbour* (KNN).

Dentre as bases de dados de domínio público, escolheu-se a contida no Sistema de Autorização Hospitalar para ser submetida às técnicas descritas, a fim de comprovar a aplicabilidade das mesmas na evidenciação dos dados anômalos para análise do auditor.

Agrupar dados é determinar diversos subconjuntos onde as distâncias entre os elementos de um subconjunto são mínimas e as distâncias entre os diversos subconjuntos são máximas. Objetiva-se agrupar dados em classes desconhecidas, utilizando medições de similaridade baseadas em distâncias entre um centro escolhido e o objeto a agrupar, com vistas a maximizar a similaridade intra-classe (os dados de um agrupamento são semelhantes entre si) e minimizar a similaridade inter-classes (os diversos grupos não guardam semelhanças) (YUFENG K., et al., 2004).

As técnicas de agrupamento rígidas determinam que um dado pertence unicamente a um *cluster*, o que não permite encontrar dados que estejam na periferia dos grupos. Relaxando tal assertiva, chegamos ao Agrupamento Nebuloso, onde todos os dados pertencem a todos os clusters com um certo grau de pertinência (XU, R., 2005). Dentre os algoritmos de agrupamentos nebulosos, o *possibilistic Fuzzy C-means* (FCM) permite atingir o objetivo deste estudo, posto que este admite a ocorrência de pertinências próximas de zero, o que caracterizaria o dado anômalo (OLIVEIRA, J.V.; PEDRYCZ, W., 2007).

Em virtude da dificuldade de tratar estruturas não lineares, adiciona-se ao algoritmo uma função kernel. Esta função mapeia os dados para um espaço de maior dimensão e permite continuar-se a usar do mesmo algoritmo (OLIVEIRA, J.V.; PEDRYCZ, W., 2007)(FILIPPONE, M. et al., 2008).

Assim, implementou-se um aplicativo com estes conceitos, que sofreu testes e avaliações, revelando a correção e precisão do mesmo.

Outra vertente escolhida foi a utilização de *SVM – One Class*, que apesar de se originar de técnicas supervisionadas de reconhecimento de padrões, não requer um conjunto de treinamento. Este método visa determinar uma fronteira ótima que separe um conjunto de dados em dois subconjuntos, um normal e outro com dados anômalos. Nesse caso, utilizou-se um aplicativo já consagrado, LIBSVM desenvolvido por CHANG, C. e LIN, C. (2001). Para facilitar seu uso desenvolveu-se uma interface, cujo teste e avaliação aquilataram sua acurácia.

Da pesquisa, verificou-se a possibilidade de estabelecer um algoritmo que indicasse os *outliers* em um conjunto de dados de forma simples, criando-se o algoritmo de Função de Similaridade Média, que semelhante aos outros utiliza a função kernel. Aqui, cada registro é caracterizado pela similaridade média deste a todos os demais, sendo considerado como dado anômalo aqueles que têm um valor pequeno. O algoritmo implementado foi avaliado e testado, resultando em um aplicativo eficaz na evidenciação dos *outliers*.

Antes de submeterem-se os dados de informações hospitalares aos algoritmos, efetuou-se a caracterização estatística da base de dados. Disso pôde-se verificar a existência de outliers à vista dos registros que superavam os limites caracterizados pelo desvio interquartilico ( $Q3 + 1,5 (Q3 - Q1)$ ). Aplicando-se a análise de componentes principais, conseguiu-se elencar as variáveis de maior importância.

Submetendo-se a base aos algoritmos, verificou-se a sua eficácia na evidenciação dos dados anômalos, cuja quantidade variou de acordo com o algoritmo e parâmetros utilizados. Fez-se dois confrontos entre os resultados, intra-métodos e inter-métodos, buscando estabelecer a concordância no apontamento dos registros anômalos. Obteve-se praticamente 100% de concordância intra-método, os dados apontados pelo método com um parâmetro em particular também foram assim rotulados, quando se utilizando um valor diverso para o mesmo parâmetro. Mais importante, é a checagem dos resultados apontados pelas diferentes técnicas. Nesse caso, confrontou-se o resultado do algoritmo de Função de Similaridade Média (FSM)

com os fornecidos pelo aplicativo de SVM – One Class e Agrupamento Nebuloso (KPCM).

Como resultado, verificou-se que há uma concordância entre os métodos FDM e KPCM em praticamente todos os casos (o índice variou entre 99,12% e 100%) e uma concordância aceitável no confronto FDM e SVM – *One Class* (índices variando entre 80,94% e 98,27%). Neste último caso, repete-se o que foi verificado na avaliação, quando o SVM – *One Class* apresentou um comportamento de rotular dados normais como anômalos, na tentativa de se obter todos os *outliers* previamente conhecidos.

Esses dados comprovam a aplicabilidade dos métodos à evidenciação do dado anômalo no intuito de auxiliar o esforço de auditoria, sendo adequados à solução do problema e exequíveis, dispensando o conhecimento prévio sobre o comportamento dos dados. Os algoritmos consagrados fornecem o suporte à eficácia do aplicativo de Função de Similaridade Média, posto que levam praticamente ao mesmo resultado, sendo que este último tem a vantagem de dispensar a introdução de parâmetros de inicialização, como a indicação de número de *clusters* no agrupamento e o valor de  $\nu$  no algoritmo SVM – *One Class*.

O aplicativo de Função de Similaridade Média, ao apresentar um resultado ordenado crescentemente, provê não apenas os dados anômalos que devem ter atenção do auditor, mas facilita a priorização daqueles que devem ser primeiramente analisados. Com isso, a análise pode ser mais extensa conforme haja maior ou menor número de analistas. Esta capacidade é inexistente no algoritmo SVM *One-Class*, onde se estabelece um subconjunto de *outliers*, sem possibilidade de se caracterizar quão mais ou menos anômalo um registro é. Na impossibilidade de se analisar todo o subconjunto, o auditor, ao selecionar uma fração, corre o risco de dispensar dados de maior valor para sua tarefa. O algoritmo KPCM permite que se ordene os registros segundo a soma de suas pertinências, igualando-se ao à FSM neste quesito, mas apresenta uma necessidade a mais de se estipular o número de grupos (e as coordenadas centrais dos clusters), o que dificultaria mais o trabalho do auditor. Assim, ressaltando que o FSM é especializado em elencar *outliers*, este algoritmo revela-se eficaz e de uso simples.

Neste trabalho, centrou-se no uso do kernel gaussiano em todos os algoritmos, sugerindo-se em futuros estudos a aplicação dos demais com o fito de se verificar um aumento na eficácia, bem como verificar se a outras formas de cálculo de distância, como Mahalanobis, afetam o resultado.

Os algoritmos Agrupamento Nebuloso (KPCM) e Função Média de Distância utilizam uma curva para evidenciar os valores da soma das pertinências e similaridade, que são utilizadas para evidenciar o ponto de separação entre os dados anômalos e normais. Tal decisão depende da decisão daquele que está executando o algoritmo e é visual. Sugere-se futuramente pesquisar uma forma mais precisa para elencar o ponto divisor.

## Referências Bibliograficas

AL HASAN, M.; CHAOJI, V.; SALEM, S.; ZAKI, M. J.; “Robust partitional clustering by outlier and density insensitive seeding”, **Pattern Recognition Letters**, Volume 30, pp. 994 – 1002, 2009

AL-ZOUBI, M. B.; “An effective Clustering-Based approach for outlier detection”, **European Journal of Scientific Research**, pp. 310-316, 2009

ANGIULLI, F.; PIZZULI, C.; “Outlier mining in large high-dimensional data sets”; **IEEE Transactions on Knowledge and Data Engineering**, Volume 17, pp. 203 – 215, 2005

BEN-HUR, A.; HORN, D.; SIEGELMANN, H.T; VAPNIL, V.; “Support vector clustering”, **Journal of Machine Learning Research**, Volume 2, pp. 125 – 137, 2001.

BENTLEY, P. ;“Evolutionary, my dear Watson: Investigating Committee-based Evolution of Fuzzy Rules for the Detection of Suspicious Insurance Claims”; **Proc. of GECCO2000**; pp. 702 – 709, 2000

BENTLEY, P.; KIM, J.; JUNG., G.; CHOI, J. ; “Fuzzy Darwinian Detection of Credit Card Fraud.”; **Proc. of 14th Annual Fall Symposium of the Korean Information Processing Society**; pp. 1 – 4; 2000

BERRUETA, L. A.; ALONSO-SANCES, R. M.; HÉBERGER, K.; “Supervised Pattern recognition in food analysis”; **Journal of Chromatography**, Volume 1158, pp. 196-214, 2007.

BITTENCOURT, F. M. R.; “Auditoria - demandas e possibilidades”; **Revista do TCU**, v. 36, n. 106, pp. 15-28, 2005

BOSE, R. P. J. C.; SRINIVASAN; “Data mining approaches to software fault diagnosis”, **15th International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications**, pp.45 – 52 , 2005

BRAUSE, R.; LANGSDORF, T.; HEPP, M.; “Neural Data Mining for Credit Card Fraud Detection”, **Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence**, pp.103-106, 1999

BUHAGIAR, A.; “Exploration and reduction of data using principal components

analysis”; **Malta Medical Journal**; volume 14; pp. 27 – 35; 2002

CALDERON, T.G.; CHEH, J.J.; “A roadmap for future neural networks research in auditing and risk assessment”, **International Journal of Accounting Information Systems**, ”, Volume 3, pp. 203-236, 2002

CAMASTRA, F.; VERRI, A.; “A novel Kernel method for clustering”, **IEEE Transactions on Pattern analysis and machine intelligence**, Volume 27, pp. 801 – 805, 2005

CAPUTO, B.; HAYMAN, E.;FRITZ, M.,; EKLUNDH, J. ;“Classifying Materials in the Real World”, **Image and Vision Computing**; 2009

CHAN, Y. H.; “Biostatistics 302. Principal component and factor analysis”; **Singapore Medical Journal**; n.45, pp. 558-566; 2004.

CHANG, C.; LIN, C; “LIBSVM : a library for support vector machines”, 2001. Aplicativo disponível em <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, acessado em 25/03/2009

CHANG, Q.; WANG, X.; LIN, Y; YEUNG, D.; “Automatic model selection for one-class SVM”; **International journal of soft computing**, Medwell Journals, pp. 307-312, 2007

CHAVES, E.; “A Subquadratic Algorithm for Cluster and Outlier Detection in Massive Metric Data”, **International Symposium on String Processing and Information Retrieval** pp. 46 - 58, 2001

CHEN, X.; AHMAD, I. S.; “Shape-Based Image Retrieval Using k-Means Clustering and Neural Networks” , **Lecture Notes in Computer Science**, pp. 893–904, 2007.

CHEN, H.; CHUNG, W.; XU, J. J.; QIN, G. W. Y.; CHAU, M.; “Crime Data Mining: A general framework and some examples”, **IEEE Computer Society**, Volume 37, pp. 50 – 56, 2004

CHEN, J.; XU,G.; “SVM-based Swift Trust Rating Model in E-commerce”; 2009 **First International Workshop on Education Technology and Computer Science**; pp. 640 - 643, 2009

CHEN, Z.; TANG, J.; FU, A. W.; “Modeling and efficient Mining of intentional Knowledge of outliers”, **Seventh International Database Engineering and**

**Applications Symposium (IDEAS'03)**, pp.44-53, 2003

CHINTALAPUDI, K. K.; KAM, M.; “A noise-resistant Fuzzy C Means Algorithm for clustering”, **Fuzzy Systems Proceedings**, Volume 2, pp.1458 – 1463, 1998

CUNHA, P. R.; BEUREN, I. M.; “Técnicas de Amostragem Utilizadas nas Empresas de Auditoria Independente Estabelecidas em Santa Catarina”, **Revista de Contabilidade e Finanças – USP**, PP. 67-86, 2006.

CUPERTINO, C. M.; MARTINEZ, A. L.; “Qualidade da Auditoria e Earnings Management: Risk Assessment através do Nível dos Accruals Discricionários”. In: **VII Encontro Brasileiro de Finanças**, 2007

DOMÍNGUES, R. A.; NANDI, A. K.; “Toward breast cancer diagnosis based on automated segmentation of masses in mammograms”; **Pattern Recognition**; Volume 42, pp. 1138-1148, 2009.

DÖRING, C.; LESOT, M.; KRUSE, R.; “Data analysis with fuzzy clustering methods”, **Computational Statistics & Data Analysis**, Volume 51, pp. 192-214, 2006

FAWCETT , T; “An introduction to ROC analysis”; **Pattern Recognition Letters**, Volume 27, PP. 861-874; 2006

FILIPPONE, M.; CAMASTRA, F., MASULLI, F.;ROVETA, S. “A survey of kernel and spectral methods for clustering” , **Pattern Recognition**, Volume 41, pp. 176-190, 2008

FILZMOSER,P.; MARONNA, R.; WERNER, M., “Outlier identification in high dimensions” , **Computational Statistics & Data Analysis**, Volume 52, pp 1694-1711, 2008

GRATERON, I. R. G; “Auditoria de Gestão: Utilização de Indicadores de Gestão no Setor Público”; **Cadernos de Estudos FIECAFI - USP**, volume 21, PP. 1-18, 1999.

GRAVES, D.; PEDRYCZ, W.; “Fuzzy C-means, Gustafson-Kessel FCM and Kernel-based FCM : A comparative study”, **Analysis and Design of Intelligent Systems using Soft Computing Techniques**, 1 ed., Springer Berlin / Heidelberg, 2007

GUNN, S. R.; **Support vector machines for classification and regression**; technical report, University of Southhampton, 1998, obtido em <http://users.ecs.soton.ac.uk/srg/publications/pdf/SVM.pdf>.

GUO, Q.; KELLY, M.; GRAHAM, C.H.; "Support vector machines for predicting distribution of Sudden Oak Death in California"; **Ecological Modeling**, Volume 182, pp. 75-90, 2005.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C.; **Análise Multivariada de Dados**; 5 ed.;S. Paulo; Bookman Companhia Editora; 2007

HAN, J.; KAMBER, M., **Data Mining, Concepts and Techniques**, 1 ed.,San Diego, Academic Press, 2001

HAO, P.; "Fuzzy one-class support vector machine", **Fuzzy Sets and Systems**, Volume 159, pp. 2317-2336 , 2008

HAWKINS, S.; WILLIAMS, G. J.; BAXTER, R. A.; CHRISTEN, P.; FETT, M. J.; HEGLAND, M.; HUANG , F.; NIELSEN, O.; SEMENOVA, T.; SMITH, A.; "Data Mining of Administrative Claims Data for Pathology Services"; **Proceedings of the 34th Hawaii International Conference on System Sciences, IEEE** , 2001

HE, J.; LAN, M.; TAN, C.; SUNG, S.; LOW, H.; "Initialization of Cluster Refinement Algorithms: a review and comparative study", **Proceedings of International Joint Conference on Neural Networks**, pp. 207-302, 2004

HE, Z.; XU, X.; DENG, S.; "Discovering cluster-based local outliers"; **Pattern Recognition Letters**, Volume 24, pp. 1641-1650, 2003

HE, Z.; XU, X.; HUANG, J.Z.; DENG, S.; "Mining class outliers:concepts, algorithms and applications in CRM"; **Experts Systems with Applications**, Volume 27, pp. 681-697, 2004

HIROTA, K.; PEDRYCZ, W.; "Fuzzy computing for Data Mining", **Proceedings of the IEEE**, Volume 87, pp.1575-1599, 1999

HODGE, V.; AUSTIN, J.; "A Survey of Outlier Detection Methodologies", **Artificial Intelligence Review**, Volume 22, pp. 85-126 , Kluwer Academic Publishers, 2004

HOFFMANN, H.; "Kernel PCA for novelty detection", **Pattern Recognition**, vol. 40, pp. 863-874, 2007.

HÖPPNER, F; KLAWONN, F.; KRUSE, R.; RUNKLER, T.; **Fuzzy cluster analysis, methods for classification, data analysis and image recognition**, 1 ed., John Wiley & Sons, Ltd., 1999.

HOREWICZ, M. C; NASCIMENTO, A. L., PERRELLA, W. J.; “Reconhecimento automático de modulação de sinais de comunicação”, Instituto Tecnológico da Aeronáutica, **IX SIGE**, 2007.

JIN, W.; TUNG, A. K. H.; HAN J.; “Mining top-n local outliers in large database” **Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining**, pp 293 - 298, 2001

KIRKOS, E. : SPATHIS, C.; MANLOPOULOS, Y.; “Data mining techniques for the detection of fraudulent financial statements”, **Expert Systems with Applications**, Volume 32, 2007, pp 995-1003

KIU, X.; CHENG, G.; WU, J. X.; “Analyzing Outliers cautiously”, **IEEE Transaction on Knowledge and Data Engineering**, vol. 14, pp. 432-437, 2002.

KNORR, E.; NG, R.; “Algorithms for mining distance-based outliers in large databases”; **International Conference in knowledge, discovery and data mining**; pp. 392-403, 1998

KOU, Y.; LU, C.; SIRWONGWATTANA, S.; HUANG, Y.; “Survey of fraud detection techniques”, **IEEE International Conference on In Networking, Sensing and Control**, Volume 2, pp. 749-754, 2004

KRISHNAPURAN, R; KELLER, J.M.; “A possibilistic approach to clustering”; **IEEE Transactions on Fuzzy Systems**, pp.98 – 110, 1993

KRISHNAPURAN, R; KELLER, J.M.; “The possibilistic C-Means algorithm: Insights and Recommendations”, **IEEE Transactions on Fuzzy Systems**, Volume 4, pp. 385-393, 1996

KROENKE, A.; SÖTHE, A.; HEIN, N.; ISHIKURA, E. R.; “Procedimentos para aplicação da amostragem estatística na Auditoria”;**XI SEMEAD**, USP, 2008.

LAM, K.; PALANEESWARAN, E.; YU, C.; “A support vector machine model for contractor prequalification”; **Automation in Construction**. Volume 18, pp.321-329 , 2008

LASKO, T.; BHAGWAT, J.; ZOU, K.; OHNO-MACHADO, L.; “The use of receiver operating characteristic curves in biomedical informatics”; **Journal of Biomedical Informatics**, Volume 38, PP. 404-415; 2005

LATECKI, L.J.; ALEKSANDAR LAZAREVIC, A.;POKRAJAC, D.; “Outlier Detection

with Kernel Density Functions”, **Lecture Notes in Computer Science**, pp. 61–75, 2007.

LESSMANN, S; VOß, S.; “A reference model for customer-centric data mining with support vector machines”; **European Journal of Operational Research**; pp. 520-530; 2009

LI, K.; TENG, G.; “Unsupervised SVM based on p-kernels for anomaly detection”; **Proceedings of the First International Conference on Innovative Computing, Information and control**, Volume 2, pp. 59 – 62, 2006.

LIN, C.; LIU, J.; HO, C.; “Anomaly detection using LIBSVM training tools”; **International Conference on Information Security and Assurance**, pp. 166 - 171, 2008.

LIOU , F.; TANG ; Y.; CHEN, J., “Detecting hospital fraud and claim abuse through diabetic outpatient services”; **Health Care Management Science**; volume 11; pp. 353-358; 2008

LITTLE, B.; JOHNSTON, W. L.; LOVELL, A. C.; REJESUS, R. M.; STEED, A. A.; “Collusion in the U.S. Crop Insurance Program : applied Data Mining”; **International Conference on Knowledge Discovery and Data Mining** , pp. 594 – 598, 2002

LU, C.; CHEN, D.; KOU, Y.; “Algorithms for spatial outlier detection”; **Proceedings of the Third IEEE International Conference on Data Mining**; pp.597 – 600, 2003

MINGOTI, S. A.; **Análise de dados através de métodos de estatística multivariada – uma abordagem aplicada**, 1 ed., Editora UFMG, 2005.

MÜLLER, K. R.; MIKA, S.; RATSCH, G.; TSUDA, K.; SCHOLKOPF, B.; “An introduction to Kernel-Based Learning Algorithms”, **IEEE Transactions on Neural Networks**, Volume 12, pp.181-201, 2001.

NETO, J. M. M.; MOITA, G. C.; “Uma introdução à análise exploratória de dados multivariados”; **Revista Química Nova**; n. 21; PP. 467 – 469; 1998.

OLIVEIRA, J.V.; PEDRYCZ, W., **Fuzzy clustering and its applications**, 1 ed., John Willey & Sons,ltd., 2007

ONODA, T.; ITO, N.; HIRONOBU, Y.; “One-class SVM based unusual condition monitoring for risk management of hydroelectric Power plants”; **Proceedings of**

**International Joint Conference on Neural Networks**; pp. 857 - 862; 2007.

ORTEGA, P. A.; FIGUEROA, P. A.; RUZ, G. A.; *A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile*; Department of Computer Science; University of Chile, 2006

PAPADIMITRIOU, S.; KITAGAWA, H.; GIBBONS, P.; FLAUTSOS, C.; *LOCI:Fast Outlier detection using Local Correlation Integral*, INTEL research, 2002

Pearson, R. K.; "Outliers in process Modeling and Identification", **IEEE Transactions on Control Systems Technology**, Volume 10, pp 55 – 63, 2002

PENG, Y.; SABATKA, A.; CHEN, Z.; KHAZANCHI, D.; KHAZANCHI, D.; SHI, Y.; "Application of Clustering Methods to Health Insurance Fraud Detection"; **International Conference on Service Systems and Service Management**, Volume 1, PP.116 - 120, 2006

PEREIRA, E. B. B.; PEREIRA, M. B.; "Um critério para o descarte de variáveis na análise de componentes principais"; **Revista Universidade Rural**, n. 1-2; PP. 1 – 7; 2004.

PÉREZ-CRUZ, F.; BOUSQUET, O.; "Kernel methods and their potential use in signal processing", **IEEE Signal Processing Magazine**, Volume 21, pp. 57 - 65 2004

PETROVSKIY, M. I.; "Outlier detection algorithms in data mining systems", **Programming and computer software**, volume 29, pp. 228-237, 2003

PHUA, C.; ALAHAKOON, D.; LEE, V. "Minority Report in fraud detection : classification of Skewed Data". **ACM SIGKDD Explorations Newsletter**, pp. 50 – 59, 2004.

PHUA, C.; LEE, V.; SMITH K.; GAYLER R.; "A comprehensive survey of Data Mining-based Fraud Detection Research"; **Artificial Intelligence Review**; 2005

PIMENTEL, E.P., FRANÇA, V, OMAR, N.; "A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização"; **XIV Simpósio Brasileiro de Informática na Educação - NCE - IM/UFRJ**, pp. 523-532, 2003

PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C.; "Curvas ROC para avaliação de classificadores"; **Revista IEEE América Latina**, pp. 215 – 222, 2008

RAMASWAMY, S.; RASTOGI, R.; SHIM, K.; "Efficient Algorithms for Mining Outliers

from Large Data Sets”; **Proceedings of the 2000 ACM SIGMOD international conference on Management of data**, pp. 427 - 438, 2000

REN, D., WANG, B., PERRIZO, W. “RDF: A Density-based Outlier Detection Method using Vertical Data Representation”, **Proceedings of the Fourth IEEE International Conference on Data Mining**, pp.503 – 506, 2004

ROSSETI, J. P., **Introdução à economia**, Editora Atlas, 1994.

SANTOS, J. S.; SANTOS, M. L. P.; OLIVEIRA, E.; “Estudo da mobilização de metais e elementos traços em ambientes aquáticos do semiárido brasileiro aplicando análises de componentes principais”; **Revista Química Nova**; n. 5; pp. 1107-1111, 2008

SCHMITT, J.; **Pré-processamento para a mineração de dados: uso da análise de componentes principais com escalonamento ótimo**; Dissertação de Mestrado; UFSC; 2005

SCHÖLKOPF, B; SMOLA, A. J.; **Learning with kernels**, 1 ed., MIT Press, 2002

SCHWERTMAN, N.C; OWENS, M.A.; ADNAN, R; “A simple more general boxplot method for identifying outliers”, **Computational Statistics & Data Analysis**, Volume 47, pp. 165-174, 2004

SHAARI, A. F.; BAKAR, A. A.; HAMDAM, A. R.; “On new approach in mining outlier”, **Proceedings of the International Conference on Electrical Engineering and Informatics**, pp. 203-206, 2007

SHALABI, L. A.; SHAABAM, Z.; “Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix” ; **Proceedings of the International Conference on Dependability of Computer Systems**, pp. 207 - 214; 2006

SHAO, H.; ZHAO, H.; CHANG G.; “ Applying data mining to detect fraud behavior in customs declaration”, **Proceedings of First International Conference on Machine Learning and Cybernetics**, Volume 3, pp.1241 – 1244, 2002

SHAW, IAN S.; SIMÕES, GODOY, M.; **Controle e modelagem Fuzzy**, 2 ed., Editora Edgard Blücher Ltda, 2001

SHEN, H.; YANG, J.; WANG, S.; LIU,X.; “Attribute weighted mercel kernel based fuzzy clustering algorithm for general non-spherical datasets”, **Soft Computing**,

Volume 10, pp. 1061-1073, 2006.

SHIN, H. J.; EOM, D.; KIM, S.; “One-class support vector machines – an application in machine fault detection and classification”; **Computer & Industrial Engineering**; Volume 48, pp. 395-408, 2005

SILVA, F. C.; “Análise ROC”, Relatório Técnico, **Instituto Nacional de Pesquisas Espaciais**, 2006

TIMM, H.; BORGEL, C.; DÖRING, C.; KRUSE, R.; “An extension to possibilistic fuzzy cluster analysis”; **Fuzzy Set and Systems**, Volume 147, pp. 3-16, 2004.

TRAN, Q.; ZHANG, Q.; LI, X.; “Evolving training method for one-class SVM”; **IEEE International Conference on Systems, Man and Cybernetics**, Volume 3, pp. 2388 – 2393, 2003.

VASCONCELOS, M.M.; HAMMERLI, I.; CAVALCANTE, M.T.L.; “Política de saúde e potencialidades de uso das tecnologias de informação”; **Rev. Saúde em Debate**, n. 61, pp. 219-235, 2002

VIAENE, S.; DERRIG, R. A.; DEDENE, G.; “A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis”; **IEEE Transactions on Knowledge and Data Engineering**, Volume 16, pp.612 – 620, 2004

VOULGARIS, Z.; MAGOULAS, G; “Extension of the K Nearest Neighbor Methods for classification problems”, **Proceedings of the 26th IASTED Conference on Artificial Intelligence and Applications**, 2008

WANG, D.; YEUNG, D. S.; TSANG, E. C. C.; “Structured One-Class classification”, **IEEE Transactions on Systems, Man, and Cybernetics**, Volume 36, pp.1283-1295, 2006

WANG, Y.; WONG, J.; MINER, A.; “Anomaly intrusion detection using one class SVM”; **Proceedings from the Fifth Annual IEEE SMC Information Assurance Workshop**, pp. 358 – 364, 2004.

WU, K.; WANG, S.; “Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space”, **Pattern Recognition**, Volume 42, pp. 710-717, 2009

Xi, J.; “Outlier Detection Algorithms in Data Mining”, **Second International Symposium on Intelligent Information Technology Application – IEEE Computer**

**Science**, pp. 94-97, 2008

XIE, Z.; WANG, S.; CHUNG, F. L.; “An enhanced possibilistic c-means clustering algorithm EPCM”, **Soft Computing - A Fusion of Foundations, Methodologies and Applications**, Volume 12, pp. 593-611, 2008.

XU, R.; WUNSCH, D.; “Survey of clustering algorithms”; **IEEE Transactions on Neural Networks**, Volume 16, pp. 645 – 678, 2005.

YANG, J.; ZHONG, N.; LIANG, P.; WANG, J.; YAO, Y.; LU, S.; “Brain Activation Detection by Neighborhood One-class SVM”; **International Conferences on Web Intelligence and Intelligent Agent Technology**; pp. 47 - 51; 2007.

YANG, M; WU, K.; “Unsupervised possibilistic clustering”; **Pattern Recognition**, Volume 39, pp. 5-21, 2006.

YUE, D.; WU, X.; WANG, Y.; LI, Y.; CHU, C.; “A Review of Data Mining-based Financial Fraud Detection Research”; **International Conference on Wireless Communications, Networking and Mobile Computing**, 2007. pp.5519 - 5522, 2007

YUFENG K., CHANG-TIEN, SIRIAT SIRWONGWATTANA, “Survey of Fraud Detection Techniques”, **IEEE International Conference on Networking, Sensing and Control**, Volume 2, pp.749 – 754, 2004

ZHANG, D.; CHEN, S; “A novel kernelized fuzzy C-means algorithm with application in medical image segmentation”; **Artificial Intelligence in medicine**, Volume 32, pp. 37-50, 2004

ZHANG, H.; WU, Q.; PU, J.; “A novel fuzzy kernel clustering algorithm for outlier detection”, **International Conference on Mechatronic and Automation**, pp. 2378-2382, 2007.

ZHANG, L.; MA, M.; LIU, X.; SUN, C.; LIU, M.; ZHOU, C.; “Differential evolution fuzzy clustering algorithm based on kernel methods”, **Lecture Notes in Computer Science**, pp. 430-435, 2006.

PROCHNOW, J. J.; “Princípio da eficiência e sua repercussão na Administração Pública”, disponível em <http://www.mt.trf1.gov.br/judice/jud5/eficiencia.htm>, 2000. Acesso em 05 de outubro de 2009

## Anexo A – Lista de Atributos do Arquivo de Autorização de Internação Hospitalar

NOME	DESCRIÇÃO
ANO_APRES	Ano de apresentação da AIH
CAR_INT	Caráter da internação, conforme a Tabela de caráter de internação.
CEP	CEP do paciente
CGC_HOSP	CGC do hospital; veja a Descrição dos arquivos do tipo CH - Cadastro de hospitais.
CID_NOTIF	CID de indicação para realização de laqueadura, conforme a Tabela da Classificação Internacional de Doenças. Não utilizado em caso de vasectomia.
COBRANCA	Motivo da cobrança, conforme a Tabela de motivos de cobrança.
COD_IDADE	Unidade de medida da idade: ignorada, dias, meses, anos
CONTRACEP1	Tipo de contraceptivo utilizado, conforme a Tabela de Contraceptivos.
CONTRACEP2	Segundo tipo de contraceptivo utilizado, conforme a Tabela de Contraceptivos.
CPF_AUD	CPF do auditor
CPF_DC	CPF do diretor clínico
DEC_APRES	Mês de apresentação da AIH
DIA_AC	Número de diárias de acompanhante
DIAG_PRINC	Diagnóstico principal, segundo CID-10 (Tabela da Classificação Internacional de Doenças)
DIAG_SECUN	Diagnóstico secundário, segundo CID-10 (Tabela da Classificação Internacional de Doenças)
DIAS_PERM	Dias de permanência; veja o Cálculo dos dias de permanência.
DT_EMIS	Data de emissão da AIH no formato aaaammdd
DT_INTER	Data de internação, no formato aaaammdd
DT_NASC	Data de nascimento do paciente no formato aaaammdd
DT_SAIDA	Data de saída, no formato aaaammdd
ESPEC	Especialidade da AIH, conforme a Tabela de especialidades.
GESTRISCO	Indicador se é gestante de risco: não é gestante de risco, é gestante de risco
IDADE	Idade, na unidade do campo COD_IDADE; veja o Cálculo da idade do paciente
IDENT	Identificação da AIH, conforme a Tabela de tipos de AIH.
INSTRU	Grau de instrução, conforme a Tabela de Grau de Instrução.
MARCA_UTI	Indica qual o tipo de UTI utilizado pelo paciente desta AIH, conforme Tabela de Tipos de UTI utilizada.
MED_RESP	CPF do médico responsável
MED_SOL	CPF do médico solicitante
MUNIC_MOV	Município onde se localiza o hospital conforme Tabela de Municípios.
MUNIC_RES	Município de residência do paciente, obtido a partir do CEP declarado de residência; veja a Descrição da Tabela de Municípios.
N_AIH	Número da AIH

<b>NOME</b>	<b>DESCRIÇÃO</b>
N_AIH_A	No caso de AIH de recém-nato que permanece após a alta da parturiente, contém o número da AIH da mãe.
N_AIH_P	É preenchido nas seguintes situações: Tendo a AIH motivo de cobrança igual a 71 (alta da parturiente com permanência do recém-nascido), contém o número da AIH do filho que permaneceu internado. Tendo a AIH motivo de cobrança de 61 a 68 (alta por reoperação), contém o número da nova AIH.
NACIONAL	Nacionalidade do paciente, conforme a Tabela de Nacionalidades.
NASC_MORT	Em caso de parto, número de nascidos mortos
NASC_VIVOS	Em caso de parto, número de nascidos vivos
NATUREZA	Natureza da relação do hospital com o SUS, conforme a Tabela de naturezas.
NUM_ENV_MO	Número do envelope
NUM_FILHOS	Número de filhos
NUM_PROC	Número do processamento, conforme a Competência dos dados e processamentos.
ORG_LOCAL	Regional do INAMPS que emitiu a AIH (em desuso)
PROC_REA	Procedimento realizado conforme Descrição da Tabela de Procedimentos.
PROC_SOL	Código do procedimento solicitado; veja a Descrição da Tabela de Procedimentos.
PRONTUARIO	Número do prontuário
QTD_CTA_MO	Quantidade de contas no envelope
SAIDA_ALTA	Em caso de parto, número de altas de neonatos
SAIDA_OBIT	Em caso de parto, número de óbitos de neonatos
SAIDA_TRAN	Em caso de parto, número de transferências de neonatos
SEXO	Sexo do paciente : Ignorado, Masculino, Feminino
TOT_PT_SP	Número de pontos de Serviços Profissionais nesta AIH.
UF_ZI	Código da UF com cuja superintendência regional o hospital mantém vinculação formal, conforme a Tabela de Unidades da Federação.
US_TOT	Valores pagos em dólares
UTI_INT_AL	Dias na UTI intermediária no mês da alta
UTI_INT_AN	Dias na UTI intermediária no mês anterior ao da alta
UTI_INT_IN	Dias de UTI intermediária no mês em que se iniciou a internação em UTI
UTI_INT_TO	Total de dias de UTI intermediária durante a internação
UTI_MES_AL	Dias na UTI no mês da alta
UTI_MES_AN	Dias na UTI no mês anterior ao da alta
UTI_MES_IN	Dias de UTI no mês em que se iniciou a internação em UTI
UTI_MES_TO	Total de dias de UTI durante a internação
VAL_ACOMP	Valor de diárias de acompanhante
VAL_MATMED	Valor de material médico
VAL_ORTP	Valor de órtese e prótese
VAL_PM	Valor pago por permanência a maior
VAL_RN	Valor de recém-nato

<b>NOME</b>	<b>DESCRIÇÃO</b>
VAL_SADT	Valor de SADT (serviços auxiliares de diagnose e terapia)
VAL_SADTSR	Valor referente a tomografias e ressonância nuclear/magnética pagas diretamente a terceiros, sem rateio
VAL_SANGUE	Valor de sangue
VAL_SH	Valor de serviços hospitalares
VAL_SP	Valor de serviços profissionais
VAL_TAXAS	Valor de taxas
VAL_TOT	Valor total da AIH: VAL_SH + VAL_SP + VAL_SADT + VAL_RN + VAL_ORTP + VAL_SANGUE + VAL_SADTSR + VAL_TRANSP
VAL_TRANSP	Valor referente a transplantes (retirada de órgãos), incluindo: taxa de sala cirúrgica (SH), retirada de órgão (SP), exames no cadáver (SADT), avaliação auditiva (SADT), exames dos transplantados (SADT)
VAL_UTI	Valor de UTI