Considerações sobre o desenvolvimento de um filtro adaptativo para validação de dados em redes de sensores (18 Outubro 2010)

Osvaldo Gogliano Sobrinho, Renata Maria Marè,
Carlos Eduardo Cugnasca, Brenda Chaves Coelho Leite

***Resumo*— Amplia-se a cada dia o uso de redes de sensores, dada a redução de custos decorrente de avanços na área de microeletrônica. Erros de leitura decorrentes de diversos motivos como, por exemplo, indução de ruído elétrico em redes cabeadas, podem levar a interpretações incorretas sobre os fenômenos observados. O uso de critérios estatísticos para validação de conjuntos de dados pode não ser adequada. Dado o caráter sazonal da variabilidade dos dados, o desenvolvimento de um filtro sensível a esta característica é altamente desejável. A utilização de tecnologias adaptativas pode levar a um modelo computacional capaz de alterar seu comportamento de maneira dinâmica, tornando possível a criação de um filtro de validação de dados com estas características.**

O trabalho apresenta considerações sobre o desenvolvimento de um filtro de validação de dados obtidos em uma rede Modbus de sensores, instalada em duas salas de aula climatizadas no Edifício de Engenharia Civil da Escola Politécnica da USP. Por tratar-se de pesquisa recém iniciada, o presente trabalho trata apenas de questões genericas levantadas até o momento para o desenvolvimento da solução.

***Palavras chave*— Tecnologia adaptativa, árvores de decisão, autômatos adaptativos, conforto ambiental, monitoramento, internet, redes de sensores.**

I. Introdução

Como parte do projeto de pesquisa intitulado "Sistema Seguro para Monitoramento Remoto da Qualidade do Ambiente Interno", patrocinado pela Fundação de Amparo à Pesquisa do Estado de São Paulo, Fapesp, dentro do programa PIPE, instalou-se uma rede de sensores padrão Modbus em duas salas de aula climatizadas do Edifício de Engenharia Civil, da Escola Politécnica da USP. Um dos objetivos da pesquisa é o monitoramento remoto de variáveis ligadas à qualidade do ar interno e seu envio a um servidor *web* remoto, onde são armazenadas e apresentadas aos usuários do sistema, por meio de um portal Internet. Assim, incorporaram-se à rede sensores analógicos (saída 0 a 5V) para medição da temperatura do ar, umidade relativa do ar e teor de CO_2.

As medições, iniciadas em setembro de 2009, são efetuadas a cada 30 segundos. No entanto, devido a problemas de latência e indisponibilidade da rede computacional do departamento, utilizada para comunicação com o servidor *web* remoto, com alguma freqüência, o intervalo entre leituras podem chegar a intervalos de 1 a 5 minutos. Em caso de queda da rede, estes intervalos podem chegar à ordem de algumas horas.

Tratando-se de uma rede cabeada, constata-se a indução de ruído elétrico em algumas medições efetuadas, levando à distorções em alguns dos valores obtidos. Faz-se necessária a filtragem dos dados de maneira a descartar estes valores. O uso de filtros estatísticos pode levar à tomada de decisões incorretas.

Discute-se neste trabalho a possibilidade de uso de um filtro baseada em uma árvore de decisão construída com o emprego de um autômato finito adaptativo.

II. Caracterização do problema

Exibe-se na figura 1 o gráfico da variação da temperatura obtida por 4 sensores e as distorções causada por dados não filtrados, que se apresentam como "picos" instantâneos de valor, que não ocorrem no ambiente monitorado.

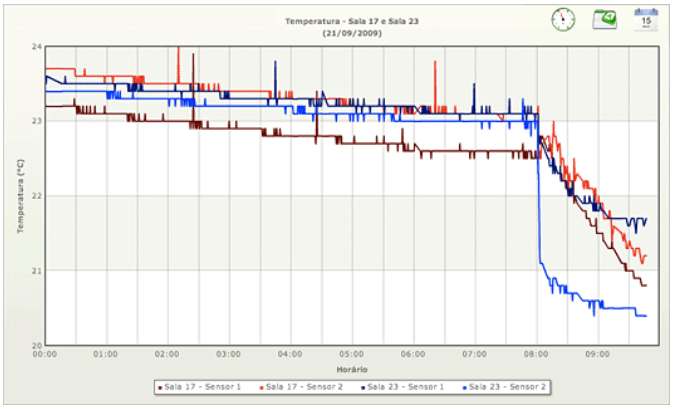

Fig. 1. Gráfico da variação de temperatura sem o uso de filtro.

A aplicação de critérios estatísticos como os de Chauvenet [1], ou Peirce [2] pode eliminar parte destes valores.

No entanto, sua utilização é conceitualmente incorreta, pois tais métodos foram desenvolvidos para a identificação de erros de leitura em conjuntos de medições distintas de uma mesma grandeza. Já no sistema de monitoramento em questão, os valores obtidos correspondem a medições efetuadas em instantes distintos, de parâmetros naturalmente variáveis no tempo.

Como é baixa probabilidade de variações significativas destes parâmetros em curtos intervalos de tempo, na prática, sua utilização pode levar a resultados aceitáveis. Observa-se na figura 2 o mesmo gráfico apos a aplicação do critério de Chauvenet, que levou à eliminação dos picos observados.

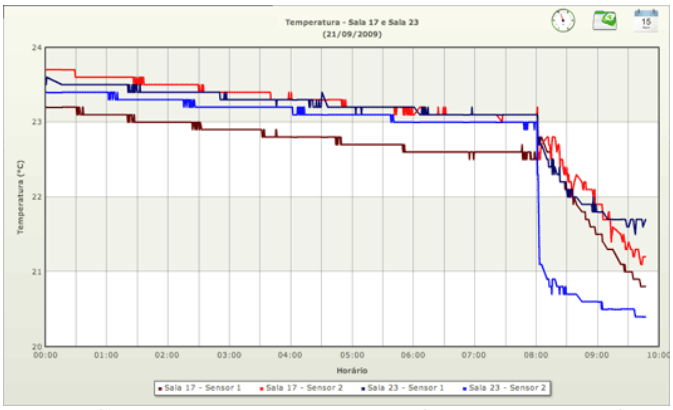

Fig. 2. Gráfico de variação da temperatura após a aplicação do critério de Chauvenet.

Outro problema que decorre do uso de filtros estatísticos é sua incapacidade de refletir as peculiaridades do ambiente monitorado. Por exemplo, nas duas salas do projeto, o sistema de climatização é ligado por volta de 08:00 de 2ª a 6ª feira, o que causa uma variação relativamente rápida da temperatura nestes momentos. Aos finais de semana, as salas não são ocupadas e o sistema de climatização permanece desligado, o que torna menor a variabilidade das condições monitoradas. Em ambas as salas, as aulas ocorrem normalmente no período da tarde, com geração de carga térmica devido à presença dos alunos e das 24 estações de trabalho que são ligadas. O efeito da presença dos alunos é notadamente marcante no teor de CO_2 medido.

Idealmente, o filtro utilizado deveria levar em conta a maior ou menor probabilidade de ocorrência destas variações em função do dia da semana e horário.

Para tentar sanar estas deficiências, imaginou-se a utilização de um filtro adaptativo que pudesse alterar dinamicamente seu critério de aceitação das medições segundo condições observadas.

III. Solução proposta

Por estar em operação desde setembro de 2009, existe uma massa de dados já coletados considerável – próximo ao final de setembro de 2010, cerca de 700.000 medições de cada parâmetro estavam registradas no banco de dados do sistema.

Imaginou-se, a princípio, a utilização de uma árvore de decisão a ser treinada com a massa de dados previamente coletada. No entanto, um requisito essencial do projeto precisa ser respeitado:

- *Possibilidade de utilização dos dados armazenados como prova em eventuais disputas jurídicas envolvendo a qualidade do ar interno dos ambientes monitorados*

Para que isso ocorra, nenhum dado pode ser descartado das medições efetuadas, pois o critério de remoção adotado certamente causará discussões sobre sua validade no âmbito jurídico. Mesmo assim, é importante que se possa visualizar os dados obtidos, por exemplo na forma de gráficos como o da figura 2, aplicando-se o filtro desenvolvido.

A conseqüência deste requisito é que o filtro não poderá ser aplicado durante a coleta de dados e sim, opcionalmente, durante a exibição dos mesmos, o que torna critico seu desempenho: note-se que para a geração do gráfico de variação de uma única grandeza (coletada a cada 30s) ao longo de um dia inteiro, cerca de 2.800 pontos por grandeza deverão ser plotados, implicando no mesmo numero de utilizações do filtro.

Com esta condição, optou-se pela utilização de uma árvore de decisão, dado seu curto tempo de resposta.

Embora a utilização do filtro deve ser feita durante a consulta aos dados, a construção da árvore pode ser feita durante a leitura dos dados já que o intervalo não menor que 30s entre medições é mais que suficiente a execução da rotina de consulta/inserção de novos ramos na árvore de decisão.

IV. Árvores de decisão

Segundo a definição de Pistori [4], "árvores de decisão (ADs) são mecanismos para a representação de funções discretas sobre múltiplas variáveis (contínuas ou discretas) com características hierárquicas que facilitam a inspeção e a utilização por seres humanos". De fato, considere-se a representação gráfica de uma árvore de decisão apresentada em exibida na figura 3.

QuickTime™ and a
decompressor
are needed to see this picture.

Fig. 3. Exemplo de árvore de decisão. Extraído de Pistori [].

A árvore apresentada representa a função

$$f : N \text{ X } L \rightarrow C$$

onde:

$$N = \{0,1,2\}$$
$$L = \{a,b\} \; e$$
$$C = \{sim, não\}$$

com os valores de C estão definidos na tabela I.

Os vértices internos da árvore representam testes efetuados sobre variáveis que conduzem, a cada possível valor a uma nova aresta representando o resultado obtido. Parte-se da raiz da árvore submetendo-se exemplos de conjuntos de atributos, chegando-se a resultados da classificação representados nas folhas da árvore.

No desenvolvimento do filtro proposto, pretende-se trabalhar com conjuntos de 4 variáveis conforme descrição no item seguinte deste trabalho.

TABELA I
DEFINIÇÃO DA FUNÇÃO $f: NXL \rightarrow C$.

N	L	f
0	a	sim
0	b	sim
1	a	não
1	b	não
2	a	sim
2	b	não

V. DISCRETIZAÇÃO DOS DADOS

Para os testes iniciais decidiu-se trabalhar com as medições efetuadas com o sensor de temperatura nº 1, instalado na Sala 17.

A seguir, definiram-se critérios para discretização dos dados. Decidiu-se pela adoção de vetores de atributos contendo quatro variáveis: dia da semana da medição, hora cheia da medição, intervalo de tempo decorrido desde a última medição e variação percentual do valor lido com relação ao último resultado. Para cada atributo consideraram-se os aspectos descritos a seguir.

A. *Dia da semana*

Para cada dia da semana, atribuiu-se um *token* $t_d \in T_d$

$$T_d = \{d0, d1, d2, d3, d4, d5, d6\}$$

para cada dia da semana, onde 'd0' corresponde à 2ª feira, 'd1' à 3ª feira e assim por diante.

B. *Hora da medição*

Decidiu-se dividir as medições em faixas baseadas em sua hora cheia, atribuindo-se a cada uma um *token* $t_h \in T_h$

$$T_h = \{h0, h1, h2, h3, ..., h22, h23\}$$

onde 'h0' corresponde ao intervalo entre 00:00 e 01:00 e 'h23' corresponde ao intervalo entre 23:00 e 24:00.

C. *Intervalo de tempo decorrido desde a última medição*

Inspecionando-se a massa de dados existentes constatou-se que 99,68% das medições ocorrem em intervalos iguais ou inferiores a cinco minutos. Desta forma, decidiu-se considerar apenas intervalos nesta faixa, atribuindo-se a cada medição um *token* $t_t \in T_t$

$$T_t = \{t1, t2, t3, t4, t5, t6\}$$

onde 't1' corresponde a intervalos iguais ou inferiores a um minuto, 't2' corresponde a intervalos iguais ou inferiores a 2 minutos e assim sucessivamente. Associou-se o *token* 't6' para quaisquer intervalos superiores a cinco minutos, situação na qual se optou pela aceitação do valor medido, ou seja, admite-se que após um intervalo de tempo superior a cinco minutos qualquer variação da leitura é admissível.

D. *Valor da medição*

Decidiu-se trabalhar com a variação percentual entre cada medição efetuada e a medição prévia. Desta maneira, o critério pode ser aplicado à medição de qualquer grandeza, o que não ocorreria se os valores fossem expressos em suas grandezas reais. Optou-se por trabalhar com intervalos percentuais de 1% já que os sensores empregados possuem uma resolução de 0,1ºC que corresponde a cerca de 0,5% na faixa usual de temperaturas observadas (20 a 30ºC).

A próxima decisão foi a escolha do numero de intervalos a considerar. Novamente, inspecionando-se a massa de dados existente, constatou-se que variações fora do intervalo -5% a +5% correspondem a apenas 0,01% da medições. Dado o equilíbrio entre variações positivas e negativas, considerou-se apenas o valor absoluta dos percentuais de variação. Assim, definiram-se 6 *tokens* $t_v \in T_v$ para os valores das medições

$$T_v = \{v0, v1, v2, v3, v4, v5, v6\}$$

onde 'v0' corresponde a variações de 0% (sem variação entre a medição atual e a anterior), 'v1' corresponde a uma variação de ±1% e assim por diante. Associou-se o *token 'v6'* a variações fora da faixa -5% a +5%. Outra decisão tomada foi o de aceitação automática de valores correspondentes a $t_v = 0$, pois é razoável a suposição que variações nulas dos parâmetros medidos sejam as mais freqüentes, dados os curtos intervalos de tempo entre medições. De fato, observando-se a massa de dados disponíveis, em 96,4% das medições efetuadas obtêm-se variações nulas.

VI. GERAÇÃO DA ÁRVORE DE DECISÃO E TREINAMENTO

A construção da árvore de decisão poderá ser feita durante a utilização do filtro. A submissão de cada vetor de atributos obtidos a um autômato adaptativo pode reconhecer uma seqüência já obtida ou, caso nova, acrescentá-la à árvore.

Para seu treinamento, árvores de decisão são submetidos a conjuntos de dados para os quais o valor da classe é conhecido o que permite seu registro. Através de algoritmos de indução da árvore de decisão é possivel a classificação de casos não presentes no conjunto de treinamento.

No caso do sistema em questão, a classificação de medições como válidas ou inválidas não pode ser feita por outro critério que não seja a probabilidade estatística de sua ocorrência. Assim, associaram-se a cada folha da árvore de decisão dois *tokens* representando o numero total de ocorrências do conjunto de dados e o numero de total de ocorrências do conjunto de dados obtido no vértice anterior. Esses *tokens* são facilmente atualizados a cada passo da rotina de geração da árvore.

Ao final, a probabilidade de ocorrência do conjunto

examinado pode ser facilmente calculada e a amostra classificada, adotando-se um valor limite para esta probabilidade.

Para a construção da árvore de decisão utilizou-se uma versão modificada do autômato classificador apresentado por Pistori e José Neto [3].

O alfabeto aceito pelo autômato modificado é definido por

$$\sum = \{t_v t_t t_d t_h\}$$

A linguagem aceita pelo autômato é:

$$t_v t_t t_d t_h$$

No autômato proposto em [3] a presença de um símbolo especial 'S' ao final da cadeia servia como seu validador, permitindo sua incorporação condicional à árvore. Em nosso caso, qualquer seqüência é reconhecida (caso previamente incorporada) ou acrescentada à árvore. Em ambos os casos os *tokens* associados aos números de ocorrência são atualizados.

VII. Resultados e pontos a serem investigados

Por apresentar resultados iniciais de uma trabalho em estágio inicial de desenvolvimento, não existem resultados a serem divulgados. Procura-se focar as primeiras averiguações em alguns pontos específicos. Dentre eles, podemos destacar:

A. Critérios de discretização

Verificar os efeitos do aumento e da diminuição dos níveis de discretização dos dados medidos, tentando balancear performance do filtro e o nível de discretização adotado. Uma menor granulação dos intervalos levará a resultados mais precisos, à custa de uma performance inferior e vice-versa.

B. Verificação da relevância de variáveis

A adoção de tokens relacionados ao dia da semana e hora de medição como atributos parece, a priori, uma idéia correta. Porém, como a diminuição do número de variáveis nos vetores de atributos leva a uma melhor performance do modelo, os efeitos de sua possível não utilização deverão ser avaliados.

C. Aprimoramento do modelo

O modelo adotado leva a resultados que podem ser incorretos. Para citar dois problemas já identificados: a primeira medição após um intervalo longo sempre é considerada correta; a medição seguinte a um ponto descartado tem probabilidade razoável de também ser indevidamente descartada, pois poderá apresentar variação semelhante (de sinal contrário) à da medição anterior.

D. Refinamento de limites para inferência estatística de aceitação dos dados

O limite adequado a ser utilizado para o descarte de medições deverá ser buscado.

Referências

[1] BOL'SHEV, L. N.; UBAIDULLAEVA, M. Chauvenet's Test in the Classical Theory of Errors. **Theory Probab. Appl**, v. 19, n. 4, p. 683-692, 1975.

[2] PEIRCE, B. Criterion for the rejection of doubtful observations. **The Astronomical Journal**, v. 2, n. 21, p. 161-163, 1852.

[3] PISTORI, H.; JOSÉ NETO, J. Adaptree - Proposta de um Algorítmo para Indução de Árvores de Decisão Baseado em Técnicas Adaptativas. In: Anais Conferência Latino Americana de Informática-CLEI, 2002, Montevideo. **Proceedings…** Montevideo: 2002.

[4] PISTORI, H. Tecnologia Adaptativa em Engenharia de Computação: Estado da Arte e Aplicações. 2003. 174 p. Tese de Doutorado - São Paulo.

Osvaldo Gogliano Sobrinho possui mestrado em Engenharia Elétrica, modalidade Sistemas Digitais pela Escola Politécnica da USP. Graduado em Engenharia Civil pela Escola Politécnica da USP (1976). Tem experiência na área de Ciência da Computação, com ênfase em Tecnologia Web e Banco de Dados. Atualmente é aluno de doutorado na Escola Politécnica da USP, na área de Sistemas Digitais e Diretor técnico da empresa Abili Assessoria Técnica Comercial e Tecnologia da Informação Ltda.

Renata Maria Marè possui mestrado em Engenharia de Construção Civil e Urbana pela Escola Politécnica da Universidade de São Paulo (2010), com ênfase em qualidade do ar de interiores em ambientes com sistemas de climatização com distribuição de ar pelo piso. Graduação em Engenharia Civil pela Escola de Engenharia Mauá (1985). É diretora comercial da Abili Assessoria Técnica Comercial e Tecnologia da Informação Ltda. desde 1996. Tem experiência nas áreas de Engenharia Civil, Design de Interiores e Paisagismo.

Carlos Eduardo Cugnasca é graduado em Engenharia de Eletricidade (1980), mestre em Engenharia Elétrica (1988) e doutor em Engenharia Elétrica (1993). É livre-docente (2002) pela Escola Politécnica da Universidade de São Paulo (EPUSP). Atualmente, é professor associado da Escola Politécnica da Universidade de São Paulo, e pesquisador do LAA - Laboratório de Automação Agrícola do PCS - Departamento de Engenharia de Computação e Sistemas Digitais da EPUSP. Tem experiência na área de Supervisão e Controle de Processos e Instrumentação, aplicadas a processos agrícolas e Agricultura de Precisão, atuando principalmente nos seguintes temas: intrumentação inteligente, sistemas embarcados em máquinas agrícolas, monitoração e controle de ambientes protegidos, redes de controle baseados nos padrões CAN, ISO11783 e LonWorks, redes de sensores sem fio e computação pervasiva. É editor da Revista Brasileira de Agroinformática (RBIAgro).

Brenda Chaves Coelho Leite possui graduação em Engenharia Civil pela Universidade Federal de Minas Gerais (1979), mestrado em Arquitetura e Urbanismo pela Universidade de São Paulo (1997) e doutorado em Engenharia Mecânica pela Universidade de São Paulo (2003). Atualmente é professor doutor da Universidade de São Paulo. Tem experiência em Conforto Térmico e Qualidade do ar nas edificações, Sistemas de Ar-Condicionado, Tecnologias de distribuição de ar pelo piso e painéis radiantes, Eficiência energética das edificações, Simulação computacional de edifícios, Dinâmica dos Fluidos Computacional, Automação e controle aplicados a sistemas de climatização, Eficiência Energética, avaliação pós-ocupação.