UNIVERSIDADE FEDERAL DO AMAZONAS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
DEPARTAMENTO DE APOIO À PESQUISA
PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

APLICAÇÃO DE MÉTODOS DE REAÇÃO À MUDANÇA EM PROBLEMAS COM AMBIENTES DINÂMICOS

BOLSISTA: Fidel Marx de Souza Guimarães, CNPq

MANAUS
2013

UNIVERSIDADE FEDERAL DO AMAZONAS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
DEPARTAMENTO DE APOIO À PESQUISA
PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

RELATÓRIO PARCIAL

PIB – E – 0175/2012

APLICAÇÃO DE MÉTODOS DE REAÇÃO À MUDANÇA EM PROBLEMAS COM AMBIENTES DINÂMICOS

Bolsista: Fidel Marx de Souza Guimarães, CNPq
Orientadora: Profa. Dra. Eulanda Miranda dos Santos

MANAUS
2013

SUMÁRIO

ÍNDICE DE FIGURAS

1. INTRODUÇÃO

O problema de classificação consiste em criar algoritmos capazes de reconhecer padrões em um determinado conjunto de informações extraídas de algum ambiente de forma a detectar e/ou prever certos eventos. No processo de classificação, considera-se que as informações permanecerão as mesmas com o passar do tempo, no entanto, a literatura indica que no mundo real, são poucos os problemas em que o ambiente é estático.

Existem diversos exemplos de problemas em ambiente chamados dinâmicos, tais como: monitoramento de fraude em cartão de crédito (KUNCHEVA, 2004), detecção de *spam* (WIDMER; KUBAT, 1996), detecção de intrusos em redes (KARNIK *et. al.*,2008), monitoramento ambiental (OZA; TUMER, 2008), dentre outros. Nesse contexto, considerar que os dados são gerados a partir de uma fonte estacionária é uma hipótese falsa, especialmente quando os dados são coletados durante um longo período de tempo, pois os conceitos não permanecem estáveis à medida que o tempo evolui (BAENA-GARCIA *et. al.*, 2006).

Áreas de pesquisa como aprendizagem de máquina, reconhecimento de padrões, estatística e mineração de dados, têm concentrado esforços na proposta de métodos de detecção e reação a mudanças. Em função da área de pesquisa, problemas com ambientes dinâmicos são chamados de ambientes não estacionários, *concept drift*, etc. Em mineração de dados, por exemplo, há problemas em que os dados são organizados na forma de fluxos, ao invés de bancos estáticos, e é bastante incomum que os conceitos e distribuições de dados permaneçam estáveis com o passar do tempo (TSYMBAL *et al*, 2008).

Para Karnick et. al.(2008), um algoritmo típico para atuar em ambientes dinâmicos precisa implementar um ou mais dos seguintes procedimentos: detecção de mudança; detecção da magnitude da mudança; aprendizagem dos novos conceitos; e esquecimento dos conceitos não mais relevantes. Esses procedimentos podem ser agrupados em dois subgrupos: (1) detecção de mudança – que envolve a detecção e a percepção da magnitude da mudança; e (2) reação à mudança – que busca manter o sistema atualizado com os conceitos mais relevantes sobre o ambiente. Este projeto concentra-se na fase de reação a mudanças, cujo objetivo é comparar diferentes métodos de reação à mudança, a fim de indicar vantagens e desvantagens dos métodos comparados. Os experimentos

envolvem o uso de bases de dados públicas que representam problemas com ambientes dinâmicos reais e sintéticos.

2. REVISÃO BIBLIOGRÁFICA

Nesta seção serão descritas algumas informações que foram incorporadas através da revisão bibliográfica, tais como o processo de classificação através de aprendizagem de máquina.

2.1. Classificadores

Classificadores são algoritmos projetados para aprender a distinguir padrões a partir de informações extraídas do ambiente. Normalmente, o processo que envolve o projeto e a implementação de um classificador é dividido em duas etapas:

1. Treinamento: o sistema aprende a classificar as diferentes classes existentes no problema;
2. Operação: o sistema identifica informações desconhecidas.

Nesse processo, portanto, assume-se que a distribuição das informações no ambiente em que o sistema de classificação irá operar permanecerá igual à distribuição representada no conjunto de treinamento, isto é, assume-se que o ambiente é estático. Entretanto, a literatura mostra que os problemas práticos do mundo real evoluem, assim como suas características, ocasionando mudanças com o passar do tempo no cenário de aplicação do sistema (KUNCHEVA, 2004). Além disso, em muitos domínios do mundo real, o conceito de interesse pode depender de algum contexto escondido, que não é fornecido explicitamente nos dados de treinamento (TSYMBAL *et al*, 2008). Portanto, estratégias devem ser utilizadas para dotar sistemas de classificação da capacidade de detectar e reagir a mudanças em ambientes dinâmicos.

Os métodos de detecção de mudança podem ser implícitos e explícitos. Nos métodos implícitos, o sistema é constantemente atualizado, ocorrendo ou não mudança. Já nos métodos explícitos, o sistema se preocupa em primeiramente perceber a mudança, para após adaptar-se ao novo conceito. Quantos aos métodos de reação à mudança, estes podem ser divididos em métodos baseados em classificadores individuais e em conjuntos de classificadores. As abordagens baseadas em classificadores individuais normalmente treinam um novo classificador sempre que uma mudança ocorre, enquanto os métodos que aplicam conjuntos de classificadores usam regras de combinação dinâmicas e heurísticas de descarte de aprendizagem para conservar o sistema invariavelmente atualizado. As estratégias investigadas neste trabalho são descritas a seguir.

2.2. Métodos de reação selecionados

Três estratégias de reação à mudança são investigadas neste trabalho: (1) classificador individual com detecção de mudança implícita; (2) Learn++.NSE – conjunto de classificadores com detecção de mudança implícita; e (3) Leave-One-Out - classificador individual com detecção explícita.

2.2.1 Classificador individual com detecção de mudança implícita

Neste método, o sistema é atualizado em períodos pré-definidos por meio do treinamento do classificador com dados atuais. Portanto, o conhecimento prévio é perdido, e o sistema mantém-se constantemente atualizado.

2.2.2 Learn++.NSE

Esse método, proposto por ELWELL e POLIKAR (2011), utiliza conjuntos de classificadores com entrada em lotes e votação majoritária ponderada. A detecção de mudança é implícita, onde a cada intervalo de tempo t, um novo classificador é gerado, chamado de t-ésima hipótese h^t. O conjunto formado por todas as hipóteses até este tempo t (incluso) é chamada de hipótese composta H^t. Com a chegada de novos dados, o algoritmo computa o erro E^t da hipótese composta (H^{t-1}) usando voto ponderado de acordo com a idade e a taxa de acerto individual de cada classificador considerando os novos dados.

2.2.3 Leave-One-Out

Neste método (PECHENIZKIY et al, 2009), um classificador é inicialmente treinado com o primeiro lote de dados. Em seguida, para cada lote subsequente, o classificador anterior é utilizado para classificar os dados por meio da estratégia *leave-one-out*, da seguinte maneira. Dado um lote com *m* dados, uma amostra é retirada do lote de dados, enquanto as *m-1* amostras restantes são usadas para testar o classificador. Esse processo é repetido *m* vezes, até que cada amostra do lote fique de fora da base de teste uma vez. Portanto, são geradas *m* taxas de erro, que em seguida são usadas para calcular a média das taxas de erro e o desvio padrão no referido lote de dados. Assume-se, para cada lote de dados que o desempenho de predição está estável através da equação:

$$E_t + S_t < E_{min} + \alpha * S_{min}$$

onde E_t e S_t representam a média de erro e o desvio padrão naquele lote , respectivamente. E_{min} e S_{min} representam o erro mínimo encontrado até o momento e o desvio padrão mínimo encontrado até o momento em todos os lotes, respectivamente, e α é um parâmetro pré-definido que representa o nível de confiança. Se a equação acima for falsa, então o algoritmo emite um alerta, indicando que existe a possibilidade de uma mudança de contexto. No entanto essa equação não representa a mudança em si, a mudança de contexto é indicada pela seguinte equação:

$$E_{min} + \alpha * S_{min} < E_t + S_t < E_{min} + \beta * S_{min}$$

onde β também é um parâmetro pré-definido que representa o nível de confiança. Se essa equação for falsa, o algoritmo reporta uma mudança e um novo classificador é gerado. Além disso, quando ocorre uma mudança de contexto S_{min} e E_{min} são redefinidos para os novos valores mínimos encontrados.

3. METODOLOGIA

Visando alcançar as metas estabelecidas por este projeto, os seguintes passos foram realizados:

3.1. Pesquisa Bibliográfica

Foi feito um levantamento sobre técnicas de detecção de mudanças e também sobre métodos de reação a mudanças. Dentre os métodos estudados foram selecionados os três métodos comparados neste trabalho, assim como as bases de dados investigadas, descritas na próxima seção.

3.2. Seleção das Bases de Dados

As 03 (três) bases de dados utilizadas são: Nebraska, CIRCLE e SINE. A base Nebraska, disponível publicamente em (ftp://ftp.ncdc.noaa.gov/pub/data/gsod), é composta por 8 atributos e 606 janelas contendo 30 amostras cada, distribuídas em 02 (duas) classes: 2 chuva e 1 não-chuva com 31% das amostras de chuva e 69% das amostras de não chuva. Trata-se de uma base de dados reais.

As bases CIRCLE e SINE são compostas por dados sintéticos e encontram-se disponíveis em (http://www.cs.bham.ac.uk/~minkull/opensource/). A base SINE apresenta mudança de conceito do tipo abrupta, enquanto a base CIRCLE apresenta mudança de conceito do tipo gradual. As duas bases são compostas por 2.500 amostras, as quais estão distribuídas em 02 (duas) classes, sendo 1.250 amostras por classe, e representam dois conceitos. A mudança de conceitos ocorre inicialmente a cada 1.000 amostras e posteriormente a cada 250 amostras.

3.3. Experimentos com as técnicas de detecção e reação à mudanças

A última etapa deste trabalho consistiu em realizar experimentos com os métodos selecionados a fim de replicar os testes realizados pelos autores. Foram utilizados para os

experimentos, os três métodos já descritos na seção anterior, bem como o uso de um classificador estático. Em todos os métodos investigados foi utilizado o classificador SVM (*Support Vector Machines*), sendo que experimentos adicionais foram realizados para a definição dos melhores parâmetros de SVM para cada base. É importante destacar que os três métodos investigados foram todos implementados neste trabalho.

4. RESULTADOS E DISCUSSÕES

Nesta seção serão descritos os resultados já alcançados por este projeto, bem como descritos os próximos passos do plano de estudo traçados a partir destes resultados.

4.1. Base Nebraska

A base Nebraska representa um ambiente cíclico, isto é, o contexto é recorrente e portanto, as classes de chuva e não chuva predominam alternadamente ao longo dos 12 meses de cada ano. Abaixo temos o gráfico do desempenho de cada um dos quarto algoritmos testados nesta base. O termo dinâmico refere-se à estratégia baseada em classificador individual com detecção implícita.

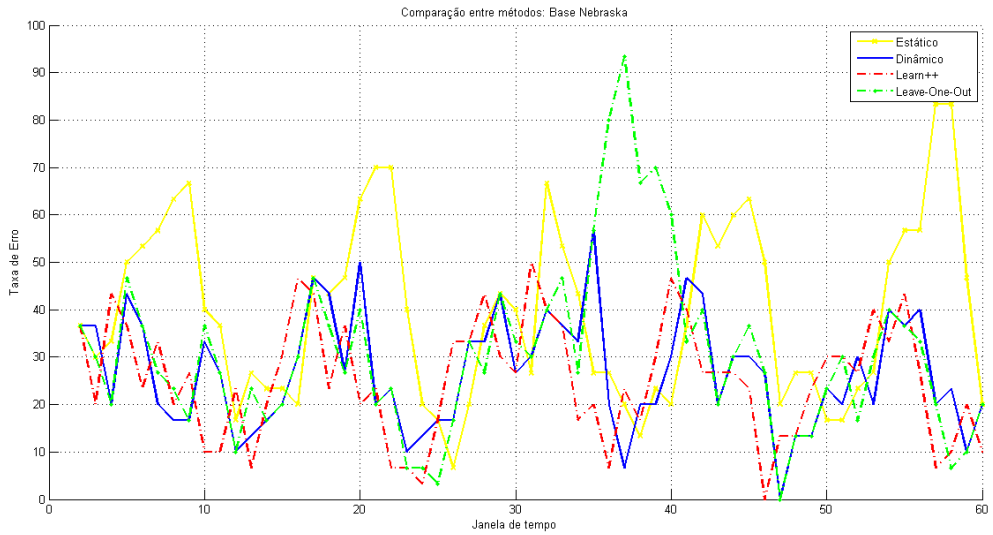

Figura 1. Desempenho dos quatro métodos na base Nebraska

O método estático, por ser treinado apenas no início do período é o menos eficaz já que "tem conhecimento" apenas de um determinado tipo de ambiente. Quanto aos demais métodos, todos obtiveram desempenho parecido, com destaque para o Learn++ que utiliza conhecimento incremental, ou seja, tem mais "conhecimento" quando comparado aos outros métodos. O *Leave-One-Out* também apresentou desempenho similar ao Learn++, porém demora um pouco mais para

se adaptar à mudança de contexto já que espera que o erro aumente consideravelmente para então re-treinar o classificador.

4.2. Base SINE

Na base SINE temos mudança de contexto do tipo abrupta. A figura 2 mostra o desempenho dos quatro algoritmos nesta base de dados.

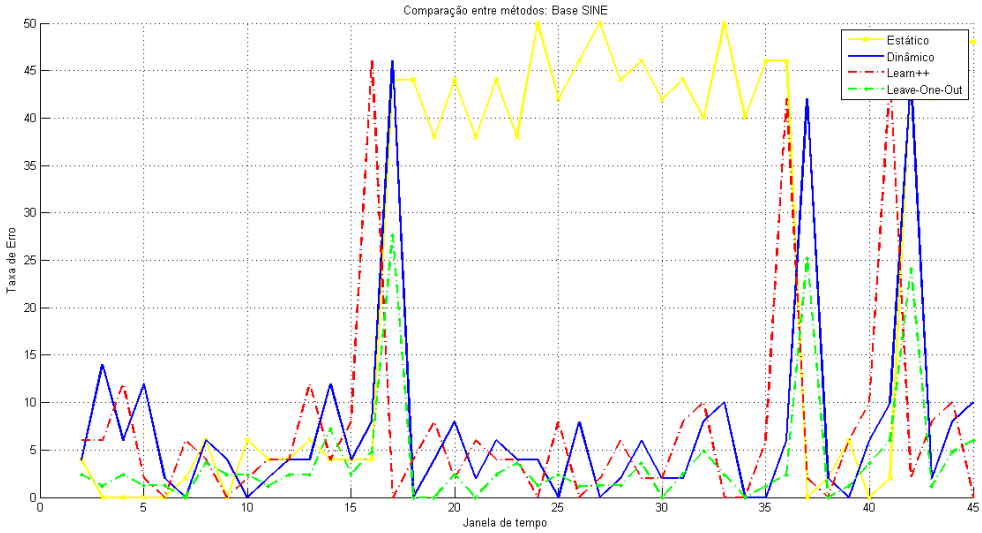

Figura 2. Desempenho dos quatro métodos na base SINE

Como pode ser visto na figura, o método estático novamente não consegue manter-se eficiente face às mudanças de contexto. Esse método apresenta taxas altas de erro devido às mudanças de conceito que ocorrem na janela 16 e na janela 41. Nesta base, destaca-se o comportamento do método *Leave-One-Out*, que obteve as menores taxas de erro dentre todos os métodos investigados, adaptando-se mais rapidamente às mudanças de contexto.

4.3. Base CIRCLE

Ao contrário da base SINE, na base CIRCLE temos mudança de contexto do tipo gradual, ou seja, o ambiente muda de forma lenta e constante, e por isso os métodos tendem a se adaptar melhor ao novo contexto. A figura 3 mostra o comportamento dos métodos nesta base.

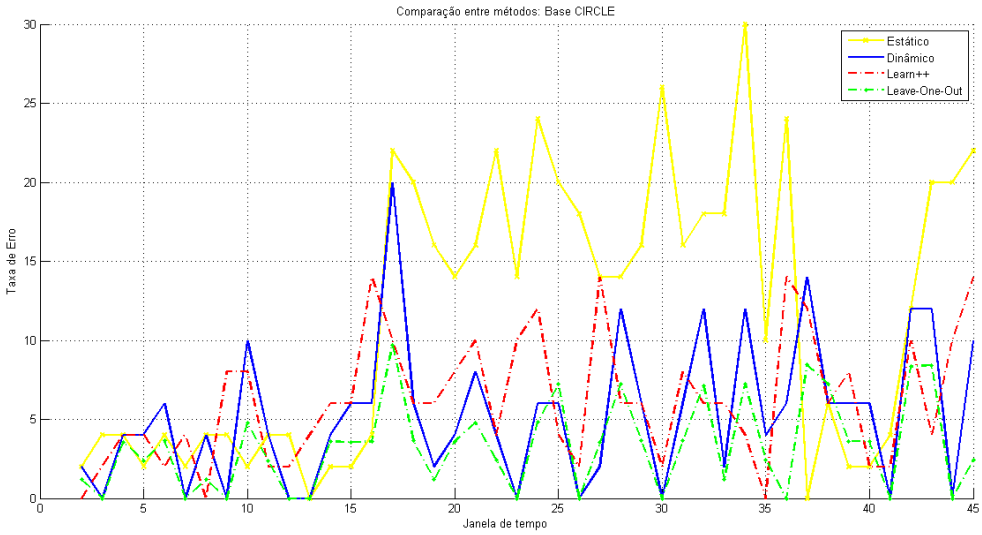

Figura 3. Desempenho dos quatro métodos da base CIRCLE

Novamente o método estático permanece com altas taxas de erro enquanto o contexto muda. Quanto aos outros métodos, destaca-se o Leave-One-Out novamente.

4.4. Conclusões

Como pode ser observado através das figuras, alguns métodos apresentam vantagem em situações específicas, enquanto outros obtiveram desempenho superior em outras. Para tentar verificar que método obteve desempenho melhor de maneira mais geral, calculou-se a média de desempenho de cada método nas três bases. A figura 4 mostra o resultado desse cálculo.

	Base Nebraska	**Base SINE**	**Base CIRCLE**
Classificador estático	60,06%	74,72%	88,55%
Classificador dinâmico	72,53%	92,64%	94,68%
Learn++.NSE	72,36%	92,90%	93,82%
Leave-One-Out	63,56%	96,15%	96,70%

Figura 4. Taxa de acerto média de cada um dos métodos nas três bases de dados.

Também foram calculados outros valores relativos ao custo computacional como quantidade de classificadores treinados. A figura 5 mostra esses dados.

	Base Nebraska	Base SINE	Base CIRCLE
Classificador estático	1	1	1
Classificador dinâmico	605	45	45
Learn++.NSE	605	45	45
Leave-One-Out	361	38	34

Figura 5. Número de classificadores treinados por cada um dos métodos nas três bases de dados.

Diante desses dados, pode-se concluir que embora os algoritmos tenham obtido resultados próximos em precisão, o método Leave-One-Out tem um custo computacional menor já que treina menos classificadores, e, portanto, faz menos uso de processamento e até mesmo de espaço em disco, já que Learn++.NSE precisa armazenar cada um dos classificadores gerados em memória. Esses pequenos detalhes podem fazer diferença se o sistema de classificação em questão envolver, por exemplo, um dispositivo móvel, o qual teria recursos limitados que precisariam ser otimizados o máximo possível.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BAENA-GARCIA, M., DEL CAMPO-ÁVILA, J., FIDALGO, R., AND BIFET, A. Early drift detection method. In *ECML PKDD 2006 Workshop on Knowledge Discovery from Data Streams,* pp. 77-86, 2006.

ELWELL, R., AND POLIKAR, R. Incremental Learning of Concept Drift in Nonstationary Environments. In *IEEE Transactions on Neural Networks*. 22 (10), pp. 1517-1531 2011.

ELWELL, R., AND POLIKAR, R. Incremental learning of variable rate concept drift. In *Multiple Classifier Systems*, pp. 142-151, 2009.

KARNICK, M., AHISKALI, M., MUHLBAIER, M., AND POLIKAR, R. Learning concept drift in nonstationary environments using an ensemble of classifiers based approach. In *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 3455 -3462, *2008.*

KUNCHEVA, L. I. Classifier ensembles for Changing Environments. In *Proceedings of Multiple Classifier System,* pp. 1-157, 2004.

OZA, N., and TUMER, K. Classifier ensembles: Select real-world applications. *Information Fusion* 9 (1), pp.4-20, 2008.

PECHENIZKIY, M., BAKKER, J., ZLIOBAITE, I., IVANNIKOV, A., KARKKAINEN, T. Online Mass Flow Prediction in CFB Boilers with Explicit Detection of Sudden Concept Drift. In ACM SIGKDD Explorations Newsletter. Volume 11 Issue 2, December 2009, 109-116.

TSYMBAL, A., PECHENIZKIY, M. AND PUURONEN, S. Dynamic Integration of Classifiers for Handling Concept Drift. *Information Fus*ion, 9 (1), pp. 56-68, 2008.

WIDMER, G., AND KUBAT, M. Learning in the presence of concept drift and hidden contexts. *Machine Learning,* 23(1), pp. 69-101, 1996.