Asian Journal of Applied Research (AJAR) 2015

E-ISSN: 2395-1702 P-ISSN: 2395-0382
Volume 01- Issue 04-, pp-01-04

Research Paper

MINING TOPIC SIGNAL FROM TEXT USING ARTIFICIAL INTELLIGENCE

N.Tamilarasi

Asst. Professor in Computer Applications
Sri Akilandeswari women's College
Vandavasi, Tamilnadu.
Ph: +917092138756
E-mail: sjarasi08@gmail.com

Abstract:

Data stream mining and sequence mining have many applications and pose challenging research problems. Typical applications such as network monitoring, web searching, telephone services and credit card purchases are characterized by the need to mine continuously massive data streams to discover up-to-date patterns, which are invaluable for timely strategic decisions. Also the data processing speed, adaptation of data, noise and mining quality are the problems to overcome. In this work, the author proposed a system which involves adaptive boosting, robust regression ensemble, sequence clustering and mining quality for better data mining.

Key words: Data stream mining, sequence mining, performance, adaptation, improving mining quality.

1. Introduction:

Today many organizations produce and consume massive data stream. Mining such data can reveal up-to-date patterns which are invaluable for timely decisions and it is different from traditional mining. First, the need of online responses requires mining to be done very fast because online systems have limited CPU power and memory. Second, the data generated by this concept is highly dynamic and the data streams are very noisy due to lack of preprocessing. A substantial amount of recent work has focused on continuous mining of data streams. Two fundamental issues are associated with continuous mining attempt (1) Performance (2) Adaptation issue. This research work proposes a novel adaptive boosting ensemble method to solve the major problems like performance and adaptation. In this, method, decision trees with a few nodes are used to achieve fast and light learning. These simple models are often weak predictive models, so the authors exploit boosting technique to improve the ensemble performance. The traditional boosting is modified to handle data streams, retaining the essential idea of dynamic sample weight assignment by eliminating the requirement of multiple pass through the data. This is then extended to handle concept drift via change detection. Change detection aims at significant changes that would cause sessions deterioration of the ensemble performance. The awareness of changes makes it possible to build an active learning system that adapts to changes promptly.

2. Related work:

As data stream mining has recently become an important research domain, concept drift is one of the central issues in stream data mining. The first systems capable of handling concept drift were STAGGER [SG86], ib3 and the PLORA family. These algorithms provide valuable insights only on small datasets; researches have not yet established the degree to which most of these AI approaches scale to large problems.

Mining quality can be improved by cleaning poor data, using more appropriate mining models or using more effective mining methods.

3. Proposed systems:

3.1 Fast & Light Stream Boosting Ensembles:

This paper evaluate boosting scheme extended with change detection, named as adaptive boosting, and compare it with weighted bagging.

Geometrically, samples are points in a 3-dimentional unit cube. The real class boundary is a sphere defined as

$$B(x) = (x_i - c_i)^{2} - r^{2} = 0$$

Robust and Adaptive stream Ensembles:

Robust Regression Ensemble method issues for adaptive learning on noisy data streams with modest resource consumption. This method assigns classifier weights in a way that maximizes the training data with the learned distribution. This weighting scheme has theoretical guarantee for adaptability and can also boost a collection of weak classifiers into a strong ensemble.

Adaptive Boosting Ensembles:

This ensemble scheme achieves adaptability by actively detecting changes and discarding the old ensemble when an alarm of change is raised. One argument is that old classifiers can be tuned to the new concept by assigning them different weights. Therefore the authors propose to learn a new ensemble from scratch when changes occur. The main challenge is to detect changes with a low false alarm rate. Two types of significant changes occurs 1.Abrupt changes2.Gradual changes.Every change detection Algorithm is a certain form of hypothesis test. To make a decision whether or not a change has occurred is to choose between two competing hypothesis: The null hypothesis H0 or the alternate hypothesis H1, corresponding to a decision of no changes or change.

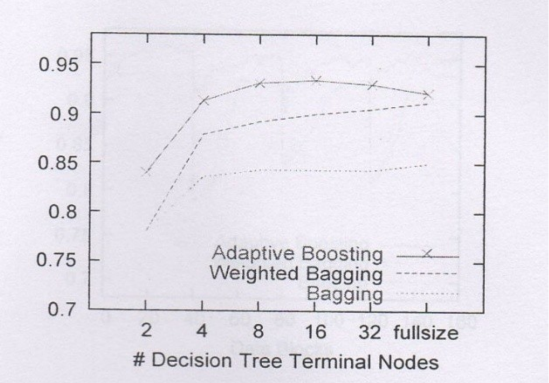
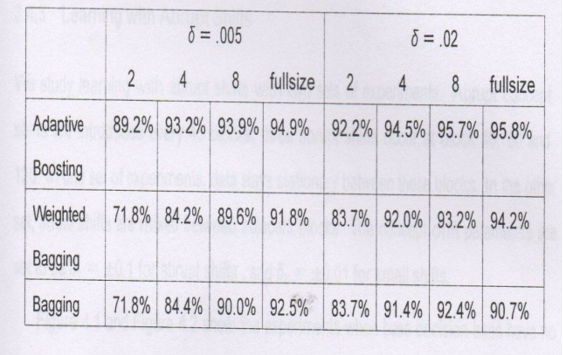


Figure1: Performance comparison of the ensembles on data with moderate gradual concept shifts



	δ = .005				δ = .02			
	2	4	8	fullsize	2	4	8	fullsize
Adaptive Boosting	89.2%	93.2%	93.9%	94.9%	92.2%	94.5%	95.7%	95.8%
Weighted Bagging	71.8%	84.2%	89.6%	91.8%	83.7%	92.0%	93.2%	94.2%
Bagging	71.8%	84.4%	90.0%	92.5%	83.7%	91.4%	92.4%	90.7%

Table 1: Performance comparison of the ensembles on data with varying levels of concept shifts.

Learning with Abrupt shifts:

Abrupt concept shifts are introduced every 40 blocks; three abrupt shifts occur at block 40, 80, and120.

Learning with Gradual shifts:

Gradual concept shifts are introduced by moving the center of the class boundary between adjacent blocks.

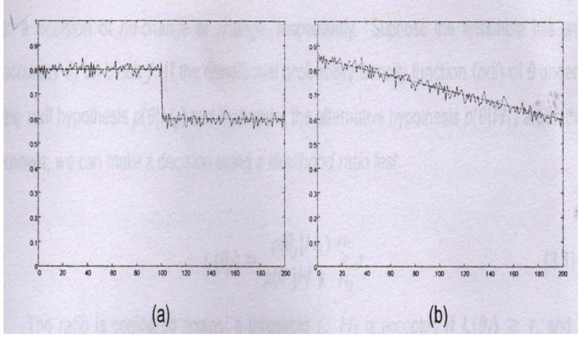


Figure 2: Two types of significant changes

a) abrupt changes b) gradual changes

Robust and Adaptive stream Ensembles

Robust regression ensemble method issues for adaptive learning on noisy data streams with modest resource consumption .This method assigns classifier weights in a way that maximizes the training data likelihood with the learned distribution .This weighting scheme has theoretical guarantee for adaptability and can also boost a collection of weak classifiers in to a strong ensemble.

Subspace Pattern based Sequence Clustering:

Clustering large datasets is a challenging data mining task. In this paper, we introduce a novel clustering model which is intuitive, capable of capturing subspace pattern similarity effectively.

The author presents a novel approach to clustering database based on the pattern similarity.

1. Comparison with previous models, the new model is intuitive for capturing subspace pattern Similarity and reduces complexity.
2. Unity pattern similarity analysis in tabular data into a single problem.
3. Present a scalable sequence used method, seqclus for clustering by subspace pattern similarity.

Mining Quality:

The usefulness of knowledge models produced by data mining depends on two issues.

1. Data Quality
2. Model adequacy

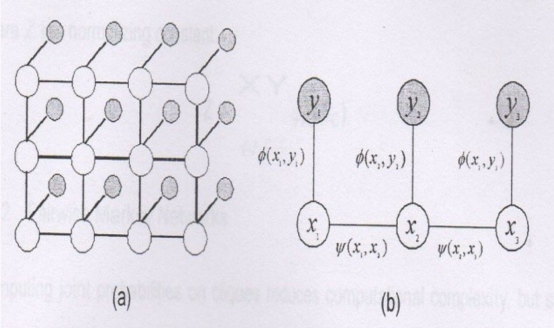
The primary contribution of this paper is that we propose a unified approach mining quality.

Markov networks:

Markov networks have been successfully applied to many problems in different fields, such as artificial intelligence image analysis, turbo decoding and condensed matter physics.

Solving a Markov network involves two phases:

1. In learning phase, a phase that builds up the graph structure of the Markov network, and learns the two types of potential functions ψ( )'s and ϕ( )'s from the training data.
2. In Inference phase, a phase that estimates the marginal posterior probabilities (or) the local maximum posterior probabilities for each random variable.



In (a) the white circles denote the random variables, and the shaded circles denote the external evidence. In (b) the potential functions are showed.

Conclusion:

This paper describes several novel algorithms in data stream mining, sequence data clustering and improving data quality in general. Adaptive boosting proposes an online boosting ensemble method that constructs a model of high accuracy and less memory consumption. To improve data quality by exploiting data interdependency using Markov random fixed modeling. Efficient inference is by belief propagation, which passes beliefs many variables so as to fill in missing values,(or) to clean the data. The author should investigated several interesting real-life applications such as cost-efficient sensor probing, protein function prediction and sequence data de-noising.

References:

1. R. Agarwal, J. Gehreand D. Gunopulos and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In proceedings of acm – sigmod international conference on management of data (SIGMOD), 1998.
2. D.Aha, D.Kilber, and M.Albert. Instance based learning algorithms. in machine learning 6(1),1991
3. S.Aji and R.McEliece. The generalized distributive law and free energy minimization. In proceedings of the 39th Annual Alteration conference on communication, control and computing, 2001.
4. C. Agarwal, C. procopiuc, j.Wolf, P.S. Yu, and J .S Park fast algorithms for projected clustering. In proceedings of ACM – SIGMOID International Conference of Management of Data (SIGMOID), 2000.
5. C. Agarwal, P.S. Yu, Finding generalized projected clusters in high dimensional spaces. In proceedings of ACM – SIGMOID International Conference of Management of Data (SIGMOID), 2000.
6. C . Agarwal, P.S. Yu, outlier detection of high dimensional data .In proceedings of ACM – SIGMOID International Conference on Management of Data (SIGMOID), 2001.
7. K. Bennett, A. Demiriz, and R. Maclin. Exploiting unlabeled data in ensemble methods. In proceedings of the 8th ACM – SIGKDD International Conference on Knowledge Discovery and data mining 2002.
8. C. Brodley and M. Friedl. Identifying and eliminating mislabeled training instances. In proceedings of the 30th National Conference on Artificial intelligence, 1996.
9. J. Blimes .A gentle tutorial on the em algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models.