POOL AND ACCURACY BASED STREAM CLASSIFICATION:

A new ensemble algorithm on data stream classification using recurring concepts detection

Mohammad Javad Hosseini, Zahra Ahmadi, Hamid Beigy
{mjhosseini,z_ahmadi}@ce.sharif.edu
beigy@sharif.edu

HaCDAIS 2011

Outline

- Proposed learning algorithm
- Evaluation and comparison of the method
- Conclusion and future works

Introduction

- We want to classify data having these properties
  - Streaming data
  - Concept drift
  - Recurring Concepts

- Previous approaches
  - The most similar method to ours is the CCP Framework [1]

Proposed Method (PASC)

- Pool of classifiers, one for each concept
- Batches of k instances

$$B_t = (x_{t,1}, x_{t,2}, \ldots, x_{t,k})$$

  - Label prediction
  - Revealing true labels

$$L_t = (l_{t,1}, l_{t,2}, \ldots, l_{t,k})$$

  - Pool update

The main framework of PASC

- PASC has three phases shown in the main loop

Input: an infinite stream of batches of instances B_t.
After classification of each instance $B_{t,i}$, its label is revealed to the algorithm.
Output: Predicted labels of instances $B_{t,i}$.

Pool = Ø; // the pool of classifiers
C = make_classifier(B_1,L_1);
RDC = new classifier(); //only used in Bayesian
//method
ac = 1; // active classifier
W_1=1;
Pool = Pool U {C};
X_1=sum_data(B_1);
RDC.update(X_1,1); //1 is the label of X_1
for j=2 to infinity **do**
    Classify B_t.
    Update Pool with B_t and L_t;
    determine active classifier (classifier weights);
end for

The main framework of PASC.

Three phases of PASC

- Phase 1: Classifying the batch
  - Active classifier method
  - Weighted classifiers method
- Phase 2: Updating the classifiers' pool
  - Bayesian method
  - Heuristic method
- Phase 3: Determining the active classifier (or classifiers' weights)
  - Needed for the next iteration

Phase 1: Classifying the Batch

Active classifier method

- Active classifier:
  - Classifier updated using the previous batch of data
    - A new classifier
    - The most relevant classifier

- If there is no concept drift
  - Active classifier would be the best choice

- When a sudden concept drift occurs
  - Performance decreases significantly

Weighted classifiers method

- A positive weight is assigned to each classifier in the beginning of processing the batch
  - According to the performance of the classifier on the previous batch

- After receiving true label of the classified instance, classifiers' weights can be updated:

$$\boxed{w'(j) = w(j) * \beta^{M(j,i)}}$$

  - $w(j)$ and $w'(j)$ are the current and new weights of the j^{th} classifier
  - β is a parameter in [0,1)
  - $M(j,i) = 0$ showing that j^{th} classifier's prediction for i^{th} instance is true and *1*, otherwise

Weighted classifiers method (Cont.)

- Using weighted majority method with these weights
  - Guaranties performance near the best classifier [2]

- But we slightly modify it to improve the efficiency
  - Uses only the classifier with the highest weight
  - Uses only a subsample of instances for updating

Phase 2: Updating the classifiers’ pool

Phase 2: Updating the classifiers' pool

- After receiving L_t, the true labels of the batch B_t:
  - Update a classifier in the batch
    - If any of the classifiers describe the batch well enough
  - Create a new classifier using this batch
    - Otherwise
- So we need to find:
  - The concept (classifier) which best describes B_t and L_t
  - A measure of the best classifier's correspondence to this batch
- This phase can be accomplished in two ways:
  - Bayesian method
  - Heuristic method

Bayesian method for Updating the classifiers' pool

- The relevance probability of the i^{th} hypothesis (h_i) to B_t and L_t is estimated as:

$$P(h_i|B_t, L_t) = \frac{P(B_t, L_t|h_i) * P(h_i)}{P(B_t, L_t)}$$

  - $P(B_t, L_t)$ is equal for all hypotheses and the prior probability of all hypotheses are the same, so the best classifier is:

$$\begin{aligned} &argmax_i P(h_i|B_t, L_t) \\ &= argmax_i\ P(B_t, L_t|h_i) \\ &= argmax_i P(L_t|B_t, h_i) * P(B_t|h_i) \end{aligned}$$

  - Conditional probability that the labels of the instances $(x_{t,1}, x_{t,2}, \ldots, x_{t,k})$ be $(l_{t,1}, l_{t,2}, \ldots, l_{t,k})$ using i^{th} hypothesis
  - Probability that the batch is produced in an environment described by the i^{th} concept

Bayesian method for Updating the classifiers' pool (Cont.)

- I.I.D condition holds for the instances in a batch
  - Size of the batch is small enough
- First term:

$$\boxed{P(L_t|B_t, h_i) = \prod_{j=1}^{j=k} P\left(l_{t,j}|x_{t,j}, h_i\right)}$$

  - Can be estimated using posterior probabilities computed by classifiers
- Second term:

$$\boxed{P(B_t|h_i) = \prod_{j=1}^{j=k} P\left(x_{t,j}|h_i\right)}$$

Bayesian method for Updating the classifiers' pool (Cont.)

- For all instances j, assume that:

$$\boxed{P(x_{t,j}|h_i) = p_i}$$

- So

$$\boxed{P(B_t|h_i) = \prod_{j=1}^{j=k} P(x_{t,j}|h_i) = {p_i}^k}$$

- To estimate p_i :
  - Construct a classifier which has a summarization of $B_{k<t}$, i.e. $X_{k<t}$ and the concept numbers of each of the batches seen so far
    - The classifier is trained incrementally to estimate the probability $P(B_t\,|\,h_i)$ by $P(X_t\,|\,h_i)$

Bayesian method for Updating the classifiers' pool (Cont.)

- To prevent underflow of the products, logarithm of the probabilities is used

$$\boxed{\begin{gathered} argmax_i P(h_i|B_t, L_t) \\ = argmax_i\, k * log\, p_i + \sum_{j=1}^{j=k} log\, P(l_{t,j}|x_{t,j}, h_i) \end{gathered}}$$

- If the pool is not full, the result of this expression is compared with a parameter θ_1
  - To decide whether a new classifier should be added.
- To improve the efficiency, only a subsample of the instances are used to estimate the best concept

Heuristic method for Updating the classifiers' pool

- Accuracy of classifiers on the new batch is used to determine the best concept

- If the pool is not full, the best accuracy is compared with a parameter θ_2.
  - To decide whether a new classifier should be added.

- To improve the efficiency, only a subsample of the instances are used to estimate the best concept

Phase 3: Determining the active classifier (or classifier weights)

Phase 3: Determining the active classifier (or classifier weights)

- Using active classifier method in phase 1
  - The active classifier should be set

- Using weighted classifiers method in phase 1
  - Classifiers' weights should be set.
  - Each classifier is tested on a subsample of the batch and weights are set by:

$$\boxed{w_0(i) = \beta^{(2^{A(i)})}}$$

  - *A(i)* is the accuracy of the i^{th} classifier

Experiments

- Datasets:
  - Emailing list Dataset [1]
  - Spam Filtering Dataset [3]
  - Hyperplane Dataset
- Updatable naïve Bayes classifier was used
- Parameter setting:
  - $\beta = 0.1$, used in weighted classifiers method
  - Maximum number of classifiers in the pool = *10*
  - Heuristic and Bayesian methods' thresholds are *0.95* and $2*\log(0.75)*m$, respectively
    - m is the subsample size used in the Bayesian method
  - Batch size = *50* for emailing list and spam filtering data sets and *500* for the hyperplane data set
- Comparisons are done in terms of accuracy, precision, recall and running time with the CCP Framework method [1]

Experiments (Cont.)

- Hyperplane data set

Classification Method	Batch assignment Method	Accuracy	Precision	Recall	Time(ms)
Active Classifier	*Bayesian*	*0.777*	*0.742*	*0.852*	*1154*
Weighted Classifiers	*Bayesian*	*0.827*	*0.781*	*0.909*	*1265*
Active Classifier	*Heuristic*	*0.760*	*0.726*	*0.838*	*1239*
Weighted Classifiers	*Heuristic*	*0.841*	*0.814*	*0.885*	*1332*
Active Classifier	*CCP (Leader follower)*	*0.758*	*0.724*	*0.836*	*1141*
Weighted Classifiers	*CCP (Leader follower)*	*0.831*	*0.807*	*0.871*	*1377*

Experiments (Cont.)

- Spam filtering data set

Classification Method	Batch assignment Method	Accuracy	Precision	Recall	Time(ms)
Active Classifier	*Bayesian*	*0.908*	*0.923*	*0.956*	*7357*
Weighted Classifiers	*Bayesian*	*0.905*	*0.938*	*0.934*	*8349*
Active Classifier	*Heuristic*	*0.908*	*0.931*	*0.947*	*3676*
Weighted Classifiers	*Heuristic*	*0.909*	*0.934*	*0.944*	*3838*
Active Classifier	*CCP (Leader follower)*	*0.915*	*0.921*	*0.969*	*3772*
Weighted Classifiers	*CCP (Leader follower)*	*0.899*	*0.932*	*0.933*	*4758*

Conclusions and future works

- The general framework used in this paper
  - Maintains a pool of classifiers
  - Updates them according to consecutive batches of data
  - Uses them to classify new batches of data
- The most similar method to ours is the CCP framework
  - Our method improves the accuracy while its parameter tuning is simpler
- Some future directions include
  - Better Management of the classifiers in the pool
  - Dynamically set the parameters according to the data sets
  - Testing more real data sets

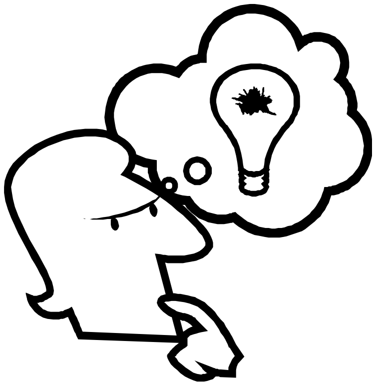
References

[1] Katakis, I., G. Tsoumakas, and I. Vlahavas, Tracking recurring contexts using ensemble classifiers: an application to email filtering. Knowledge and Information Systems, 2009. **22**(3): pp. 371-391.

[2] Freund, Y. and R.E. Schapire. Game theory, on-line prediction and boosting. in Proceedings of the ninth annual conference on Computational learning theory. 1996: ACM.

[3] Machine Learning & Knowledge Discovery Group. accessed on June 2011; Available from: http://mlkd.csd.auth.gr/concept_drift.html.


THANK
YOU



You can contact us for any questions

{mjhosseini,z_ahmadi}@ce.sharif.edu

beigy@sharif.edu

Experiments (Cont.)

Classification Method	Batch assignment Method	Accuracy	Precision	Recall	Time(ms)
Active Classifier	*Bayesian*	0.759	0.735	0.762	1862
Weighted Classifiers	*Bayesian*	0.832	0.828	0.812	2411
Active Classifier	*Heuristic*	0.748	0.718	0.768	1490
Weighted Classifiers	*Heuristic*	**0.828**	**0.811**	**0.830**	**1440**
Active Classifier	*CCP (Leader follower)*	0.771	0.748	0.775	1311
Weighted Classifiers	*CCP (Leader follower)*	0.816	0.791	0.828	1860

Experiments (Cont.)

- Comparison of methods' accuracies
  - Using weighted classifiers method leads to much better accuracies (about *8%*) in data sets with sudden concept drift
  - Our batch assignment methods (Bayesian and heuristic) have results similar to the CCP framework method while having simpler parameter tuning
- Comparison of methods' run times
  - Active classifier method
    - Heuristic and CCP methods are similar
    - Bayesian takes the most time
  - Weighted classifier method
    - Heuristic takes the least and Bayesian takes the most time