Recursive Ensemble Approach for Incremental Learning of Non-Stationary Imbalanced Data

Pradnya A. Jain
ME Scholar
Dr. D.Y.Patil School of Engineering and Technology, Lohegaon,pune – 412 105,India

Roshani Raut (Ade)
Assistant Professor
Dr. D.Y.Patil School of Engineering and Technology, Lohegaon,pune – 412 105,India

P.R.Deshmukh, Ph.D
Professor
Sipna shikshan prasarak mandal's, Amaravati - 444701

ABSTRACT
Learning non-stationary data stream is much difficult as many real world data mining applications involve learning from imbalanced data sets. Imbalance dataset consist of data having minority and majority classes. Classifiers have high productivity accuracy on majority classes and Low productivity accuracy on minority classes. Imbalanced class partition over data stream demands a technique to intensify the underrepresented class concepts for increased overall performance. To alleviate the challenges brought by these problems, this paper propose the recursive ensemble approach (REA). This approach reduces problem of imbalance data by learning minority and majority instances arrived at incremental time. In Practical analysis REA results are compare with Synthetic Minority Over-sampling Technique (SMOTE) and predicted results proves that REA gives better performance as compare to SMOTE on synthetic and real time datasets.

Keywords
Class Imbalance, Incremental Learning, Non-Stationary, REA, SMOTE.

1. INTRODUCTION
Informally, concept drift refers to a change in the class (concept) definitions over time, and therefore a change in the distributions from which the data for these concepts are drawn. An environment from which such data is obtained is a non-stationary environment. Concept drift frequently occurs in the real world. For example, people's preferences for products change. The factors that determine a successful stock change with the economy. When factory conditions change, the process for validating a product changes as well. Many times the cause of change is hidden, leaving it to be inferred from the classifications themselves. Algorithms that track concept drift must be able to identify a change in the target concept without direct knowledge of the underlying shift in distribution. With the continuous expansion of data availability in many large-scale, complex, and networked systems, such as surveillance, security, Internet, and finance, it becomes critical to advance the fundamental understanding of knowledge discovery and analysis from raw data to support decision-making processes. Although existing knowledge discovery and data engineering techniques have shown great success in many real-world applications, the problem of learning from imbalanced data (the imbalanced learning problem) is a relatively new challenge that has attracted growing attention from both academia and industry. The imbalanced learning problem is concerned with the performance of learning algorithms in the presence of underrepresented data and severe class distribution skews. Due to the inherent complex characteristics of imbalanced data sets, learning from such data requires new understandings, principles, algorithms, and tools to transform vast amounts of raw data efficiently into information and knowledge representation.

A number of solutions to the class-imbalance problem were previously proposed both at the data and algorithmic levels. At the data level, these solutions include many different forms of re-sampling such as random oversampling with replacement, random undersampling, directed oversampling (in which no new examples are created, but the choice of samples to replace is informed rather than random), directed undersampling (where, again, the choice of examples to eliminate is informed), oversampling with informed generation of new samples, and combinations of the above techniques. At the algorithmic level, solutions include adjusting the costs of the various classes so as to counter the class imbalance, adjusting the probabilistic estimate at the tree leaf (when working with decision trees), adjusting the decision threshold, and recognition-based (i.e., learning from one class) rather than discrimination-based (two class) learning. Mixture-of-experts approaches (combining methods) have been also used to handle class-imbalance problems. These methods combine the results of many classifiers; each usually induced after over-sampling or under-sampling the data with different over/under-sampling rates.

In this paper, we propose a Recursive Ensemble Approach (REA) in an effort to provide a solution for handling imbalanced data streams of nonstationary class concepts. Different from (Gao et al. 2007), REA takes a similar step as SERA to incorporate part of previous minority class examples into the current training data chunk. However in lieu of limiting the availability of hypotheses on the current training data chunk as in SERA as well as in literature (Gao et al. 2007), REA combines all hypotheses built over time in a dynamically weighted manner to make predictions on the testing data set. The proposed REA framework in this work is mainly motivated by our recent approach of MuSeRA (Chen and He 2010). Specifically, in this paper we investigate a different strategy of estimating the similarity between previous minority class examples and the current minority class set. Furthermore, based on the success of SERA (Chen and He 2009) and MuSeRA (Chen and He 2010), in this work we significantly extend simulations of REA to both synthetic benchmarks and real-world data sets. We also further design various simulations to test the robustness of REA under different parameter Evolving Systems settings. Such empirical results together with the theoretical analysis provide a more

comprehensive justification of the effectiveness of the proposed REA framework.

2. LITERATURE SURVEY

In this section we are discussing the different algorithms presented for concept drift as well as class imbalance problems.

2.1 Related Work on Imbalance Data Stream

Please Learning from data stream has been featured in many practical applications such as network trafficmonitoring and credit fraud identification. Generally speaking, data stream is a sequence of unbounded, real-time data items with a very high rate that can be read only once by an application. The restriction placed by the end of this definition is also called one-pass constraint (Aggarwal 2007), which is also claimed by other literature (Muhlbaier et al. 2009). It has been

flourished for quite a few years for the studies of learning from data stream. In Angelov and Zhou (2006), an approach to real-time generation f fuzzy rule-based systems of eXtended Takagi-Sugeno (xTS) type from data streams was proposed, which applies incremental clustering procedure to generate clusters to form fuzzy rule based systems. (Georgieva and Filev 2009) proposed the Gustafson-Kessel Algorithm for incremental clustering of data stream. It applies adaptive-distance metric to identify clusters with different shape and orientation. As a follow-up, (Filev and Georgieva 2010) extended Gustafson-Kessel Algorithm to enable real-time clustering of data stream. In Dovzan and Skrjanc (2010), a recursive version of the fuzzy identification method and predictive functional model is proposed to the control of a nonlinear, time-varying process. Incapability of storing all data into memory for learning as done by traditional approaches has yet been the solely challenge data stream has presented to the community. As what it sounds to be, concept drift, also recognized as time-evolving nature, suggests it is undesirable yet inevitable that most of the time class concepts evolve as data stream forwards. This property combined with virtually unbounded volume of data stream accounts for the so-called ‘‘stability-plasticity’’ dilemma. One may be trapped in an endless loop of pondering either reserving just the most recent knowledge to battle against concept drift or keeping track of knowledge as much as possible in avoidance of ‘‘catastrophic forgetting’’. With regards to this, many works have been recorded to strike a balance between two ends of the ‘‘stability-plasticity’’ dilemma. Marked as an effort of adapting ensemble approach to time-evolving data stream, SEA maintains an ensemble pool of C4.5 hypotheses with a fixed size, each of which is built upon a data chunk with unique time stamp. When the request of inserting a new hypothesis is made but ensemble pool has been fully occupied, some criterion is introduced to evaluate whether the new hypothesis is qualified enough to be accommodated at the expense of popping an existing hypothesis therein. The potential problem for this approach is the choice of granularity for cross validation. As straightforward as it can be, finer granularity would more accurately provide the desirable portion of old data. However increasing performance comes with extra overhead. When granularity is tuned fine enough to the scale of single example, cross validation would degenerate itself into a brute force method, which may exhibit intractability for applications sensitive of speed. Despite the popularity of data stream study, learning from nonstationary data stream with skewed class distribution is a relatively uncharted area, of which the difficulty resides itself in the context. In static context, the counterpart of this problem is recognized as ‘‘imbalanced learning’’ which corresponds to domains where certain types of data distribution over-dominates the instance space compared to other data distribution (He and Garcia 2009). It is a recently emerged area and has attracted significantly growing attention in community (Hong et al. 2007; Masnadi-Shirazi and Vasconcelos 2007). However the same story does not come to the same problem in the context of data stream, where the number of solutions is rather limited. Those on record include (Gao et al. 2007) which accommodates all previous minority class examples into the current training data set to compensate skewed class distribution, upon which an ensemble of hypotheses is built. In lieu of this aggressive accommodation mechanism, our previous work SERA (Chen and He 2009) chooses a portion of previous minority class examples into the current training data chunk based on their similarity. Accumulation of previous minority class examples is of limited volume due to skewed class distribution. Therefore, it should not be considered as violation of one-pass constraint.

3. IMPLEMENTATION

3.1 SMOTE

This approach is inspired by a technique that proved successful in handwritten character recognition (Ha & Bunke, 1997). They created extra training data by performing certain operations on real data. In their case, operations like rotation and skew were natural ways to perturb the training data. We generate synthetic examples in a less application-specific manner, by operating in “feature space” rather than “data space”. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. Our implementation currently uses five nearest neighbors. Synthetic samples are generated in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features.

3.2 REA

Recursive Ensemble Approach (REA) in a shot to produce an answer for handling unbalanced information streams of non-stationary category concepts. Completely different from (Gao et al. 2007), REA takes a similar step as SERA to include a part of previous minority category examples into the present coaching information chunk. but instead of limiting the provision of hypotheses on the present coaching information chunk as in SERA as well as in literature (Gao et al. 2007), REA combines all hypotheses designed over time during a dynamically weighted manner to create predictions on the testing information set. Over-sampling approaches, like SMOTE/SMOTEBoost, and data Boost-IM (Hong et al. 2007) produce artificial minority category instances primarily based upon existing minority category examples to balance skew class distribution. REA conjointly seeks to amplify the amount of minority category examples within the current coaching information chunk. But rather than making artificial minority category instances, REA collects minority examples from previous coaching data chunks over time and by selection accommodates those with high similarity with this

minority category set into the current coaching information chunk. Therefore we use REA for reduce imbalance data and rebalance with Learn++ concept drift.

3.3 Dataset

We will use Electricity pricing Dataset, the dataset provides time and demand fluctuations within the worth of electricity in New South Wales, Australia. We tend to use the day, period, authority (New South Wales) electricity demand, VIC (Victoria) electricity demand and therefore the scheduled electricity transfer as the features. And we will also use Whether Dataset. This dataset may be a set of the National Oceanic and Atmospheric Administration knowledge that we have a tendency to initial processed and free as an idea drift dataset. While the knowledge set contains weather data from many locations round the world, we have a tendency to selected Offutt Air Force Base in Bellevue, Nebraska as a result of this location contained over fifty years value of information, providing not solely alternating seasonal changes, however conjointly presumably long-run global climate change. Daily measurements were taken for a spread of options such as temperature, pressure, visibility, and wind speed. We have a tendency to selected eight options and set the classification task to predict whether rain precipitation was determined on daily. We have a tendency to use a test-then-train strategy for evaluating the algorithms to cast this as a prediction drawback, instead of a pure classification task.

4. PRACTICAL RESULTS AND ENVIRONMENT

In this section we representing the practical environment, such as dataset used, and metrics computed. In our current study, we adopted the classification and regression tree (CART) as the base classifier. The strategy of making CART output likelihoods that the input instance should belong to any class with is twofold. (1) The leaf node that the instance under testing falls in is located; (2) inside that leaf, proportions of training examples belonging to each class are calculated as the likelihood of the instance under testing for each class. The reason of choosing CART as the base learner is that it can provide desired trade-off between speed and performance. Base learners such as logistic regression or decision stump are not strong enough to efficiently learn knowledge from data chunks with unnatural class distribution. Other base learners such as neural networks of multi-layer perceptron (MLP) and support vector machines (SVMs) are obviously strong enough to effectively learn from streamed data chunks.

4.1 Input Dataset:

We have used following datasets to match the performance of each strategies in terms of exactness, recall and accuracy rates.

Following table shows the number of different Imbalance streaming dataset used.

Table 1. Input Dataset

Number	Dataset Name
1.	Electricity Pricing Dataset
2.	Whether dataset

4.2 Hardware and Software Used

Hardware Configuration

- Processor - Pentium –IV
- Speed - 1.1 Ghz
- RA - 256 MB(min)
- Hard Disk - 20 GB
- Key Board - Standard Windows Keyboard
- Monitor - SVGA

Software Configuration

- Operating System - Windows XP/7/8
- Programming Language - Java
- Tool - Netbeans.

4.3 Mathematical Implementation Input Dataset:

We used above dataset given in table 1 as input.

Setup phase:

All experiments begin at time $t = 1$

Input for REA:

- Minority class ratio of training data stream γ.
- Post-balance ratio f.
- Training data chunk S_t.
- Testing data set T_t.
- k -parameter for k-nearest neighbors selective accommodation mechanism
- $\mathcal{G}$ -Dataset consist of minority samples befor current time t

Process for REA:

For t=1,2,…

1. $|P_t|/|S_t| = \gamma$, where P_t is minority class samples
2. Check for post rario.
   If $f <$ (t-1) $\times$ γ
   -Check $\mathcal{G}$ to make $S_t's$ minority class ratio f
   -Add $\mathcal{G}$ to current training data chunk S_t to make $S_t^{'} = \{S_t, \mathcal{G}\}$
   Else
   -Compute k-nearest neighbors for each minority example in $\mathcal{G}$ to get δ_j
   -Sort $\{\delta_j\}$ and associate to M
   -Include M into Curent data $S_t^{'} = \{S_t, M\}$
   End if
3. Build hypothesis, $H_{t-1} = \{H_t, h_t\}$
4. For each hypothesis Derive weight as:

$$e_i = \frac{1}{S_t} \sum_{(x_j, y_j) \in S_t} (1 - f_{y_j}^i(x_j))^2$$

$$w_i = \log(\frac{1}{e_i})$$

Where, $f_{y_j}^i(x_j)$ is output probability

5. Append $P_t$into $\mathcal{G}$ i.e. $\mathcal{G} = \{\mathcal{G}, P_t\}$

Output:

Final hypothesis $h^{(t)}$ for T_t.

$$h_{final}^{(t)}(x_j^{'}) = \arg \max_{y^{'} \epsilon Y} \sum_{i=1}^{t} w_i \times h_i(x_j^{'}, y^{'})$$

4.4 Metrics Computed

As we use two different approach for class imbalance and the minority class data and the majority class data belong to positive and negative classes, we computed true positive (TP), false positive (FP), true negative (TN), or false negative (FN) etc.

4.5 Results of Practical Work

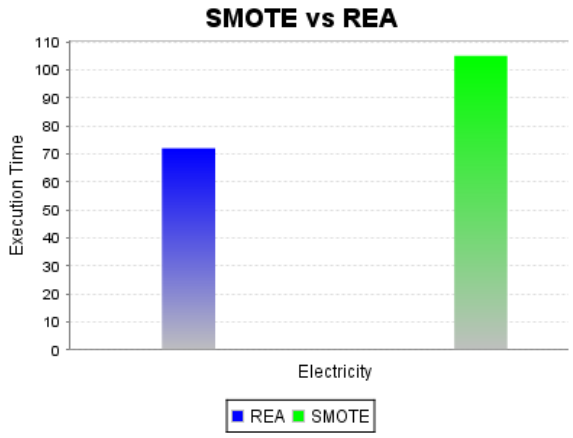

Figure 3: execution time comparative graph for electricity dataset

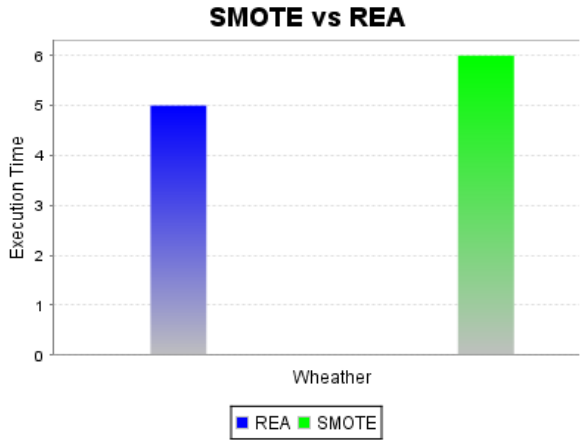

Figure 4: execution time comparative graph for wheather dataset

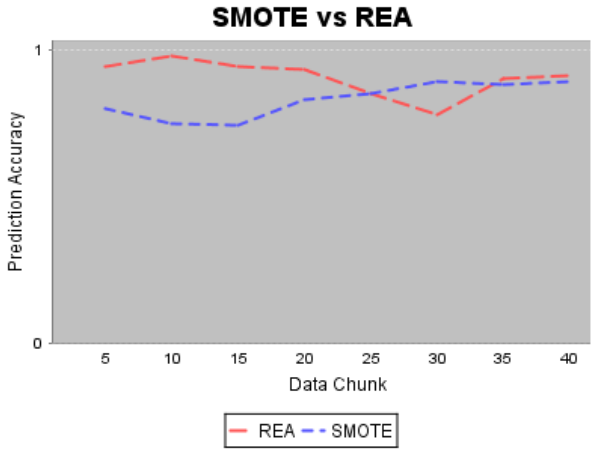

Figure 4: Overall Accuracy predicted

6. CONCLUSION AND FUTURE SCOPE

In this paper we have presented the comparison between two methods used for removal of Non-Stationary Environments from Live Imbalanced Data Streaming with aim of improving the performance. In this paper we presented the architecture of REA algorithm to improve the performance of existing SMOTE approach. The main of this paper was to present the comparative study of methods for class imbalanced problems. The recent method SMOTE is presented for providing the efficient solution to above problem in [1]. The key idea of this approach is to estimate the similarity between previous minority class examples and the current minority class set based on k-nearest neighbor, and then selectively accumulate a certain amount of previous minority class examples into the current data chunk to compensate the skewed class distribution. REA does not take significantly more time than other comparative algorithms to conduct training on data stream, which makes it to be a qualified candidate for real-time online learning system.

7. REFERENCES

[1] Gregory Ditzler, *Student Member IEEE*, and Robi Polikar, *Senior Member IEEE, "*Incremental Learning of Concept Drift from Streaming Imbalanced Data", 2012.

[2] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research 16 (2002) 321–357

[3] Sheng Chen • Haibo He," Towards incremental learning of nonstationary imbalanced data stream: a multiple selectively recursive approach", August 2010.

[4] Ha, T. M., & Bunke, H. (1997). Off-line, Handwritten Numeral Recognition by Perturbation Method. *Pattern Analysis and Machine Intelligence*, *19/5*, 535–539.

[5] Aggarwal C (2007) Data streams: models and algorithms. Springer, New York Angelov P, Zhou X (2006) Evolving fuzzy systems from data streams in real-time. In: IEEE symposium on evolving fuzzy systems. IEEE Press, Ambelside, pp 29–35

[6] Chen S, He H (2009) Sera: selectively recursive approach towards nonstationary imbalanced stream data mining. IEEE-INNSENNS international joint conference on Neural Networks, pp 522–529

[7] Chen S, He H (2010) Musera: multiple selectively recursive approach towards imbalanced stream data mining. In: Proceedings of world conference computational intelligence

[8] Dovzan D, Skrjanc I (2010) Predictive functional control based on an adaptive fuzzy model of a hybrid semi-batch reactor. Control Eng Practise 18(8):979–989

[9] Filev D, Georgieva O (2010) An extended version of the gustafsonkessel algorithm for evolving data stream clustering. In: Angelov P, Filev D, Kasabov N (eds) Evolving intelligent systems: methodology and applications. IEEE Press Series on Computational Intelligence, Wiley, pp 273–300

[10] Gao J, Fan W, Han J (2007) On appropriate assumptions to mine data streams: analysis and practice. In: Proceedings of international conference data mining, Washington, DC, USA, pp 143–152

[11] Gao J, Fan W, Han J, Yu PS (2007) A general framework for mining concept-drifting streams with skewed distribution. In: Proceedings of international conference SIAM

[12] Georgieva O, Filev D (2009) Gustafson-kessel algorithm for evolving data stream clustering. In: Proceedings of international conference computer systems and technologies for PhD students in computing

[13] He H, Chen S (2008) Imorl: Incremental multiple-object recognition and localization. IEEE Trans Neural Netw 19(10):1727–1738

[14] He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowledge Data Eng 21(9):1263–1284

[15] Hong X, Chen S, Harris CJ (2007) A kernel-based two-class classifier for imbalanced data-sets. IEEE Trans Neural Netw 18(1):28–41

[16] Masnadi-Shirazi, Vasconcelos N (2007) Asymmetric boosting. In: Proceedings of international conference machine learning

[17] Muhlbaier MD, Topalis A, Polikar R (2009) Learn??.nc: Combining ensemble of classifiers with dynamically weighted consult-andvote for efficient incremental learning of new classes. IEEE Trans Neural Netw 20(1):152–168