ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

An Ensemble Classification Framework to Evolving Data Streams

A.Sasinth[1], M.Kavitha[2], Dr.S.Karthik[3]

PG Scholar, SNS College of Technology, Coimbatore, India[1]

Assistant Professor, SNS College of Technology, Coimbatore, India [2]

Professor and Dean, SNS College of Technology, Coimbatore, India [3]

ABSTRACT: **Data stream classification poses many challenges to the data mining community. In this paper, we address four such major challenges, namely, infinite length, concept-drift, concept-evolution, and feature-evolution. Since a data stream is theoretically infinite in length, it is impractical to store and use all the historical data for training. Concept-drift is a common phenomenon in data streams, which occurs as a result of changes in the underlying concepts. Concept-evolution occurs as a result of new classes evolving in the stream. Feature-evolution is a frequently occurring process in many streams, such as text streams, in which new features (i.e., words or phrases) appear as the stream progresses. Most existing data stream classification techniques address only the first two challenges, and ignore the latter two. In this paper, we propose an ensemble classification framework, where each classifier is equipped with a novel class detector, to address concept-drift and concept-evolution. To address feature-evolution, we propose a feature set homogenization technique. We also enhance the novel class detection module by making it more adaptive to the evolving. Stream and enabling it to detect more than one novel class at a time. Comparison with state-of-the-art data stream classification techniques establishes the effectiveness of the proposed approach.**

I. INTRODUCTION

Data mining is the process of extracting or mining knowledge from large amount of data. It is an analytic process designed to explore large amounts of data in search of consistent patterns and systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. It can be viewed as a result of natural evolution of information in development of functionalities such as data collection, database creation, data management, data analysis. It is the process where intelligent methods are applied in order to extract data patterns from databases, data warehouses, or other information repositories. The data mining is a step in the knowledge discovery process. The data mining step interacts with a user or a knowledge base. There are different data repositories on which mining can be performed. The major data repositories are relational databases, transactional databases, time-series databases, text databases, heterogeneous databases, and spatial databases.

The data mining concept can be classified into two types: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

II. CLUSTERING

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. Clustering is also called data segmentation in some applications. A cluster of training data is unknown. Clustering is the assignment of objects into groups called clusters.

Data cleaning is the process of detecting and correcting in accurate records from a record set, table or data base. After cleaning, a data set will be consistent with other similar data sets in the systems. The inconsistencies detected or removed are originally caused by user entry errors or corruption in transmission. Data cleaning differs from data validation means data is rejected from the system at entry and is performed at entry time rather than batches of data. Data cleaning includes the process of parsing, data transformation, duplicate elimination and statistical methods. Record deduplication is the process of identifying and removing duplicate entries in a repository. It is also referred as data cleaning, record linkage and matching.

III. TYPES OF CLUSTERING

3.1 WELL-SEPARATED CLUSTERS

A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.

3.2 CENTER-BASED CLUSTERS

A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster.

3.3 CONTIGUOUS CLUSTERS

A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

3.4 DENSITY-BASED CLUSTERS

A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. Used when the clusters are irregular or intertwined, and when noise and outliers are present.

IV. TRAINING PHASE

A k-NN-based classifier is trained with the training data. Rather than storing the raw training data, K clusters are built using a semi-supervised K-means clustering, and the cluster summaries (mentioned as pseudopoints) of each cluster are saved. These pseudopoints constitute the classification model. The summary contains the centroid, radius, and frequencies of data points belonging to each class. The radius of a pseudopoint is equal to the distance between the centroid and the farthest data point in the cluster.

Each pseudopoint corresponds to a "hypersphere" in the feature space with a corresponding centroid and radius. The decision boundary of a model Mi is the union of the feature spaces encompassed by all pseudopoints h 2 Mi. The decision boundary of the ensemble M is the union of the decision boundaries of all models Mi 2 M.

4.1 CLASSIFICATION AND NOVEL CLASS DETECTION

Each instance in the most recent unlabeled chunk is first examined by the ensemble of models to see if it is outside the decision boundary of the ensemble. If it is inside the decision boundary, then it is classified normally (i.e., using majority voting) using the ensemble of models. Otherwise, it is declared as an F-outlier, or filtered outlier.

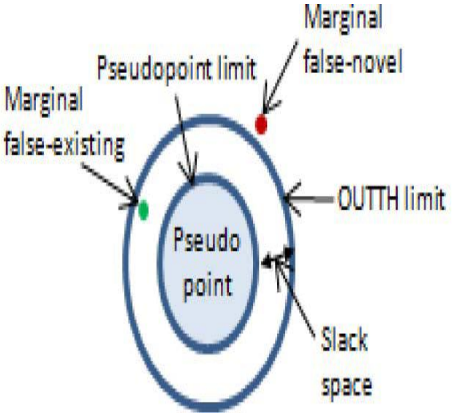
4.2 FEATURE SPACE CONVERSION

It is obvious that the data streams that do not have any fixed feature space (such as text stream) will have different feature spaces for different models in the ensemble, since different sets of features would likely be selected for different chunks. There are three possible alternatives: 1) Lossy fixed conversion (or Lossy-F conversion in short), 2) Lossy local conversion (or Lossy-L conversion in short), and 3) Lossless homogenizing conversion (or Lossless in short).

4.3 NOVEL CLASS DETECTION:

The novel class detection technique in three ways, which are 1) outlier detection using adaptive threshold, 2) novel class detection using Gini coefficient, and 3) simultaneous multiple novel class detection.

Outlier Detection Using Adaptive Threshold a slack space beyond the surface of each hypersphere. If any test instance falls within this slack space, then it is considered as existing class. This slack space is defined by a threshold, which is referred to as OUTTH.



Simultaneous Multiple Novel Class Detection The main idea in detecting multiple novel classes is to construct a graph, and identify the connected components in the graph. The number of connected components determines the number of novel classes.

V. CONCLUSION

Clustering merely tries to identify groups of similar items within the data and report these back to the user. This falls into the class of "unsupervised learning" techniques, in contrast to "supervised learning", which requires a training set of data to be made available which is tagged with the appropriate class identifier?

REFERENCES

1. Aggarwal.C.C(2009), "On Classification and Segmentation of Massive Audio Data Streams," Knowledge and Information System, vol. 20, pp. 137-156,
2. Bifet.A, Holmes.G, fahringer.B.P, Kirkby.R, and Gavalda.R(2009)`, "New Ensemble Methods for Evolving Data Streams," Proc. ACMSIGKDD 15th Int'l Conf. Knowledge Discovery and Data Mining,pp. 139-148,
3. Chen.S, Wang.H, Zhou.S, and Yu.P(2008), "Stop Chasing Trends: Discovering High Order Models in Evolving Data," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 923-932, .
4. Fan.W(2004), "Systematic Data Selection to Mine Concept-Drifting DataStreams," Proc. ACM SIGKDD 10th Int'l Conf. Knowledge Discoveryand Data Mining, pp. 128-137, .
5. Hulten.G, Spencer.L, and Domingos.P(2001), "Mining Time-Changing Data Streams,"Proc. ACM SIGKDD Seventh Int'l Conf. Knowledge Discovery and Data Mining, pp. 97-106,. MASUD ET AL.: classification and adaptive novel class detection of feature-evolving data streams 1495 Fig. 5. Scalability of MCM with number of dimensions.
6. Katakis.I, Tsoumakas.G, and Vlahavas.I(2006), "Dynamic Feature Space and IncrementalFeature Selection for the Classification of Textual Data Streams," Proc. Int'l Workshop Knowledge Discovery from Data Streams (ECML/PKDD), pp. 102-116,
7. Katakis.I, Tsoumakas.G, and Vlahavas.I(2010), "Tracking Recurring Contexts Using Ensemble Classifiers: An Application to Email Filtering," Knowledge and Information Systems, vol. 22, pp. 371-391,
8. Kolter.J and Maloof.M(2005), "Using Additive Expert Ensembles to Cope with Concept Drift," Proc. 22nd Int'l Conf. Machine Learning (ICML), pp. 449-456,
9. Lewis.D.D, Yang.Y, Rose.P, and Li.F(2004), "Rcv1: A New Benchmark Collection for Text Categorization Research," J. Machine Learning Research, vol. 5, pp. 361-397,.
10. Li.X, Yu P.S, Liu.B, and Ng.K(2009), "Positive Unlabeled Learning for Data Stream Classification," Proc. Ninth SIAM Int'l Conf. Data Mining (SDM), pp. 257-268, .