Pattern Recognition - Lecture 11
Class Imbalanced Problem & Active Learning

Dr. Dewan Md. Farid

Associate Professor, Department of Computer Science & Engineering
United International University, Bangladesh

May 11, 2018

Class Imbalanced Problem

Active Learning

Data Pre-processing

Feature Selection

Data Balancing Methods

Classification of multi-class imbalanced data is a difficult task, as real data sets are noisy, high dimensional, small sample size that results overfitting and overlapping of classes..

- Traditional machine learning algorithms are very successful with classifying majority class instances compare to the minority class instances.
- The conventional data balancing methods alter the original data distribution, so they might suffer from overfitting or drop some potential information.

We proposed a new method for dealing with multi-class imbalanced data based on clustering and selecting most informative instances from the majority classes.

Classifying Imbalanced Data

Machine learning algorithms successfully classify majority class instances, but misclassify the minority class instances in many high-dimensional data sets.
Following methods are used for class imbalance problems:

1. Sampling methods
   - Under-sampling
   - Over-sampling
2. Cost-sensitive learning methods (difficult to get the accurate misclassification cost)
3. Ensemble methods
   - Bagging
   - Boosting

Data Balancing Method

- Initially, we cluster the majority class instances into several clusters.
- Find the most informative instances in each cluster. The informative instances are close to the center of cluster and border of cluster.
- Then several data sets are created using these clusters with most informative instances by combining the instances of minority classes.
- Every data set should have almost equal number of minority-majority classes instances.
- Finally, multiple classifiers are trained using these data sets. The voting technique is used to classify the existing/ new instances.

Data Balancing Method (con.)

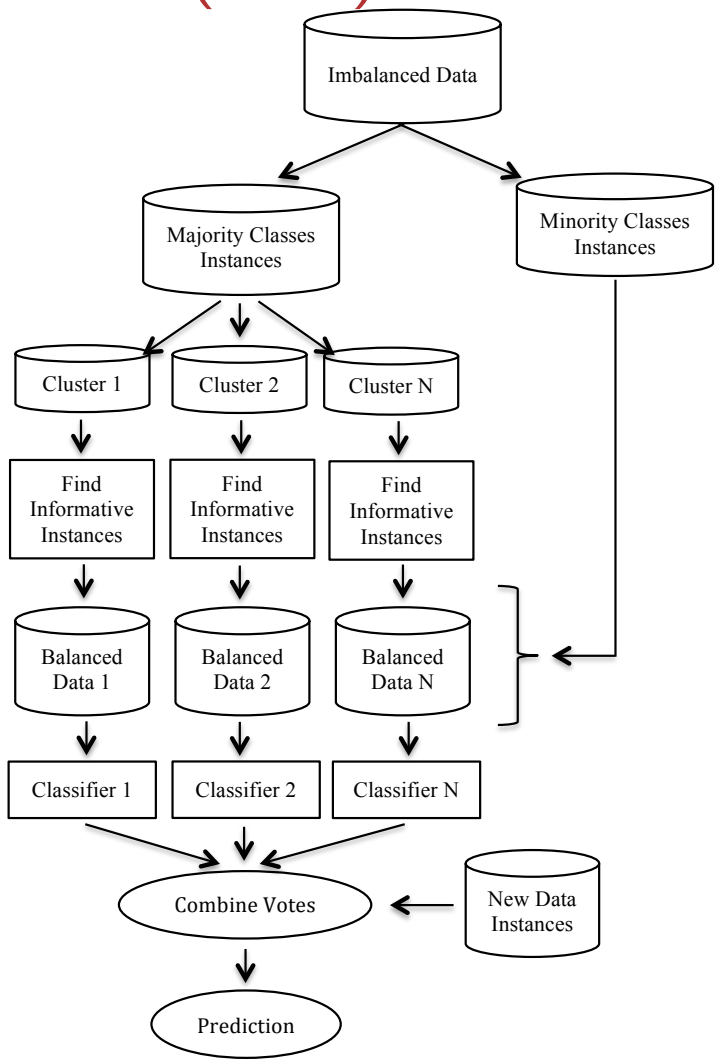

Figure: Data balancing method.

Active Learning

It achieves high accuracy using the number of instances to learn a concept can often be much lower than the number required in typical supervised learning.

- It interactively queries a user/ expert for class labels of unlabeled instances.
- The objective is to train a classifier using as few labeled instances as possible by selecting the most informative instances.

Let the data, D contains both set of labeled data, D_L and set of unlabeled data, D_U. Initially, a model, M^* trains using D_L. Then a **querying function** uses to select unlabeled instances, $X_U \in D_U$ and requests a user for labeling, $X_U \rightarrow X_L$. After X_L is added to D_L and train M^* again. The process repeats until the user is satisfied.

Active Learning (con.)

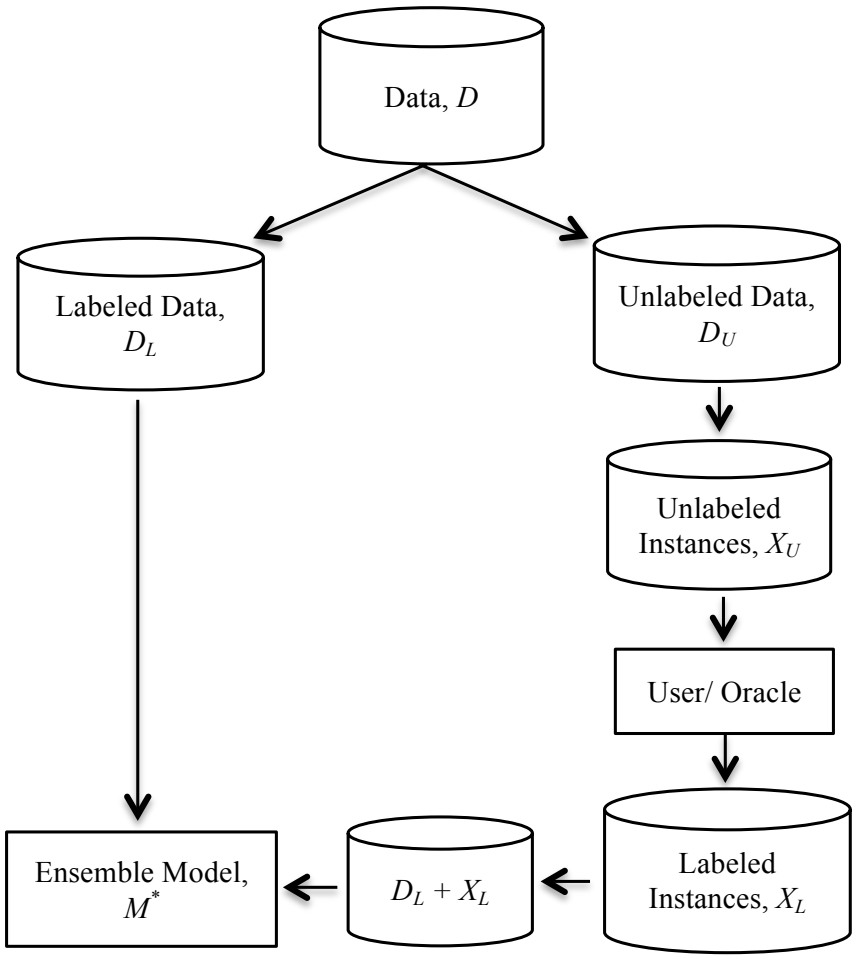

Figure: Active learning process.

An Active Learning Method

The naïve Bayes (NB) classifier and clustering are used to find the most informative instances for labeling as part of active learning. The unlabeled instances are selected for labeling using the following two strategies:

- Instances close to **centers of clusters** and **borders of clusters**.
- If the **posterior probabilities** of instances are **equal/ very close**.

Data Pre-processing

It transforms raw data into an understandable format, which includes several techniques:

- **Data cleaning** is the process of dealing with missing values.
- **Data integration** merges data from different multiple sources into a coherent data store like data warehouse or integrate metadata.
- **Data transformation** includes the followings: (a) normalisation, (b) aggregation, (c) generalisation, and (d) feature construction.
- **Data reduction** obtains a reduced representation of data set (eliminating redundant features/ instances).
- **Data discretisation** involves the reduction of a number of values of a continuous feature by dividing the range of feature intervals.

Feature Selection

It is the process of selecting a subset of relevant features from a total original features in data.
Mainly the following three reasons are used for feature selection:

- Simplification of models
- Shorter training times
- Reducing overfitting

In big data, features may contain false correlations and the information they add is contained in other features. We can apply an unsupervised feature selection approach based on measuring similarities between features by maximum information compression index. We can quantify the information loss in feature selection with entropy measure technique.

*** THANK YOU ***