DATA MINING
INTERIM REPORT

Analysis of Structure-based Techniques for Preprocessing Induced Hypergraphs of Massive Datasets for Transductive Learning

By

S K Ramnandan (CS11B061)
Varun Gangal (CS11B038)

Department of Computer Science and Engineering
IIT Madras

Submitted On
March 26, 2014

1 Abstract

The proliferation of the Internet has resulted in a rapid increase in the production of published information. This has led to the emergence of a large class of machine learning problems where the challenge has shifted from scarcity of data to lack of reliably labelled data. Such tasks are not amenable to inductive learning techniques. In these and many other situations, transductive learning is good alternative. The aim of this project is to design a mechanism for preprocessing the available labelled data using a suitable graphical representation of the dataset. The efficacy of different structural techniques in filtering useful labels is to be studied. Also, an attempt will be made to extend this system to enable active learning.

2 Motivation

Transductive learning is a method of reasoning from specific training cases to specific test cases, in contrast to inductive learning where the aim is to obtain a generic functional mapping from an input space to an output space.

A simple situation in which transductive learning techniques would be preferred is a problem for which the training data and many test instances are available a-priori but the number of labelled training instances given is far less than the number of available unlabelled test instances. As most induction methods rely on the availability of an extensive, representative training set, an inductive learning function is unlikely to be able to capture all the facets of the input space. Further, in such a problem, as the "test set" is known beforehand, a transductive algorithm, which by its nature would be tailored to the specific instances available, would possibly construct a richer model than that which a generic (inductive) approach would create. Hence, the transductive algorithm could be expected to perform better.

Another situation in which transductive learning could be useful is a binary classification problem in which inputs naturally tend to cluster into two groups, based in their class labels.

The next question that arises is how much of the available labels are to be used. It is possible that some of the given labels are noisy. The information obtained from unlabelled points could be used to discard such labels. Also, some labelled points may be outliers or present along cluster boundaries and hence may not be truly representative of the labels of the points nearest to them.

A graphical model has been chosen to represent the dataset for preprocessing. One of the main benefits of such a representation is that it aids visualization of the data. Further, many properties of graphs such as degree, connectedness or clusters provide rich information that would otherwise remain latent.

3 Problem Definition

Given a dataset of labelled and unlabelled instances for a classification task, the first step is to discretize the continuous-valued features and induce a hypergraph on the data. The structural properties of this hypergraph are to be used to filter or rank labelled points based on their relevance. Following this, graph diffusion is to be used to label the remaining points.

4 Plan of Action

1. ***Construction of hypergraph*** - The first step is to obtain the graphical representation of the dataset. Every data point will correspond to a node in the hypergraph and all nodes that share an attribute value (or attribute value level in the case of continuous-valued attributes) will be connected by a hyperedge. Thus, for every attribute, the number of hyperedges corresponds to the number of discrete values or levels possible for the attribute.

2. ***Preprocessing of labelled points*** - In this phase, different structural properties of the hypergraph are to be examined for their efficacy in filtering out noisy or non-representative labels. Ideally, the aim is to divide hypergraphs into different classes such that for a class of hypergraphs, a particular property, or group of properties is an optimal choice for preprocessing the labelled points.

3. ***Complete the classification*** - The final phase consists of using either graph diffusion on the hypergraph or by clustering the dataset directly and important labels (as weighted in step 2) within clusters to classify the unlabelled points.

5 Evaluation

In order to test the strength of the techniques used, they will be applied on fully labelled datasets. A subset of the points will be chosen to act as the labelled training set and the remaining points will be classified using the above method. A comparison of the results of the algorithm with the available labels is to be used to determine the quality of the technique in general and the choice of the structural properties of the induced hypergraph used for filtering.

6 Challenges

- The choice of values for discretization of continuous-valued features is likely to affect the induced hypergraph structure

- The nature of the near-neighbour graph cannot be determined a-priori from the dataset. So it is possible that many datasets will have to be explored to obtain different classes of graphs.

7 Dataset

http://snap.stanford.edu/

References

[1] Dengyong Zhouy, Jiayuan Huangz and Bernhard Scholkopfx - Learning with Hypergraphs: Clustering,Classification, and Embedding

[2] Cecile Bothorel and Mohamed Bouklit - An algorithm for detecting communities in folksonomy hypergraphs

[3] Evo Busseniers - General Centrality in a hypergraph

[4] Li Pu and Boi Faltings - Hypergraph Learning with Hyperedge Expansion