Wrocław University of Technology

Various applications of restricted Boltzmann machines for bad quality training data

Maciej Zięba

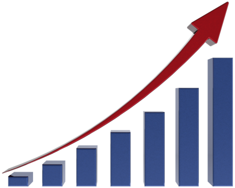
Wroclaw University of Technology

20.06.2014

Motivation

Big data - 7 dimensions[1]

- **Volume**: size of data.
- **Velocity**: speed, displacement of data.
- **Variety**: diversity of data.
- **Viscosity**: measures the resistance to flow in the volume of data.
- **Virality**: measures how fast data is distributed unique and shared between nodes in a network (e.g. the Internet).
- **Veracity**: trust and quality of the data.
- **Value**: what is the added value that Big Data should bring?

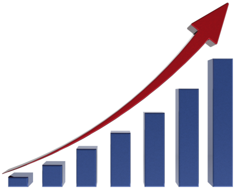


[1] According to ATOS company

Motivation

Big data - 7 dimensions[1]

- **Volume**: size of data.
- **Velocity**: speed, displacement of data.
- **Variety**: diversity of data.
- **Viscosity**: measures the resistance to flow in the volume of data.
- **Virality**: measures how fast data is distributed unique and shared between nodes in a network (e.g. the Internet).
- **Veracity**: trust and quality of the data.
- **Value**: what is the added value that Big Data should bring?



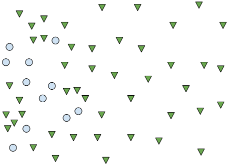
[1] According to ATOS company

Veracity of Data

Typical problems with data - training context

- **Imbalanced data problem**. One class dominates another in the training data.
- **Noisy labels problem**. Some of the examples in training data contain incorrectly assigned labels.
- **Missing values issue**. Values of some features are unknown.
- **Unstructured data**. The data is represented in unprocessed form: images, videos, documents, XML structures.
- **Semi-supervised data**. Some portion of training data is unlabelled.

Example of imbalanced data

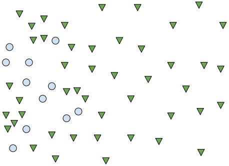


Veracity of Data

Typical problems with data - training context

- **Imbalanced data problem**. One class dominates another in the training data.
- **Noisy labels problem**. Some of the examples in training data contain incorrectly assigned labels.
- **Missing values issue**. Values of some features are unknown.
- **Unstructured data**. The data is represented in unprocessed form: images, videos, documents, XML structures.
- **Semi-supervised data**. Some portion of training data is unlabelled.

Example of imbalanced data



Methods

Restricted Boltzmann Machines (RBM)

- **RBM** is a **bipartie Markov Random Field** with **visible** and **hidden** units.

- The **joint distribution** of visible and hidden units is the **Gibbs** distribution:

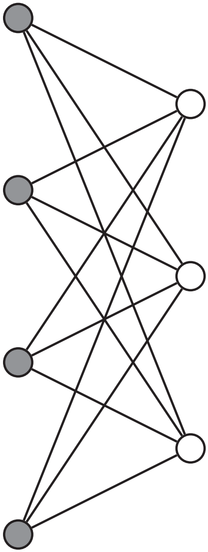
$$p(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta}) = \frac{1}{Z}\exp\big(-E(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta})\big)$$

- For **binary visible** $\mathbf{x} \in \{0,1\}^D$ and **hidden** units $\mathbf{h} \in \{0,1\}^M$ th energy function is as follows:

$$E(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta}) = -\mathbf{x}^\top \mathbf{W}\mathbf{h} - \mathbf{b}^\top \mathbf{x} - \mathbf{c}^\top \mathbf{h},$$

- Because of **no visible to visible**, or **hidden to hidden** connection we have:

$$p(x_i = 1|\mathbf{h}, \mathbf{W}, \mathbf{b}) = \text{sigm}\big(\mathbf{W}_{i\cdot}\mathbf{h} + b_i\big)$$
$$p(h_j = 1|\mathbf{x}, \mathbf{W}, \mathbf{c}) = \text{sigm}\big((\mathbf{W}_{\cdot j})^\top \mathbf{x} + c_j\big)$$



Methods

RBM for imbalanced data

- Train the model on **examples** from **minority** class by application of **MLL** (scaled):

$$\frac{1}{N}\log\left(p(\mathcal{X}_{n=1}^{N}|\boldsymbol{\theta})\right) = \frac{1}{N}\sum_{n=1}^{N}\log\Big(\sum_{\mathbf{h}} p(\mathbf{x}_n, \mathbf{h}|\boldsymbol{\theta})\Big)$$

- Generate artificial examples $\bar{\mathcal{X}}_{m=1}^{M}$ using Synthetic Oversampling TEchnique (**SMOTE**).

- For each of the newly created example $\mathbf{x}_m$ apply **Gibbs** sampling:

$$\mathbf{h}_m \sim p(\mathbf{h}|\bar{\mathbf{x}}_m, \theta)$$
$$\tilde{\mathbf{x}}_m \sim p(\mathbf{x}|\mathbf{h}_m, \theta)$$

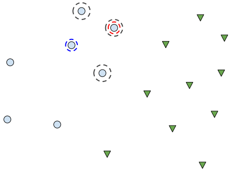
- Label newly created example $\tilde{\mathbf{x}}_m$ and store in training data.

Methods

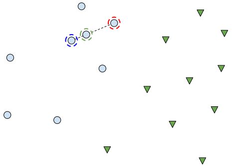
RBM for imbalanced data - example

SMOTE procedure:

A



B

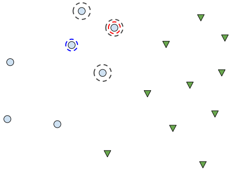


Methods

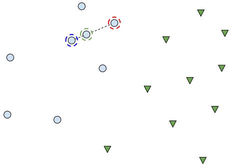
RBM for imbalanced data - example

SMOTE procedure:

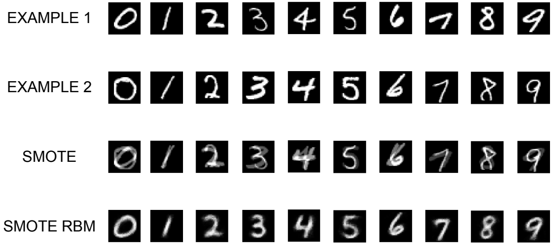
A



B



Generating artificial examples on MNIST data:



RBM for other raw data issues

- Problem of **missing values**.
  - **RBM** is trained for **each of the classes** separately.
  - **Gibbs** sampling is applied to **uncover** unknown values.
  - RBM models are iteratively **updated** while new training **example is completed**.
- Problem of **noisy labels**.
  - **RBM** is trained for **each of the classes** separately.
  - Each of the trained models is used as an **oracle** to detect **uncorrected labelled data**.
  - **Reconstruction** error is used to determine **unlabelled examples**.
- Problem of **unstructured data**.
  - **RBM** is used as domain-independent **feature extractor** that transforms raw data into **hidden units**.