Corroboration

Alban Galland

GEMO Seminar

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE

Goal

- Context : Set of sources stating facts
- Problem : find which facts are true and which facts are false
- Which information to use?
  - Functional dependencies
  - Number of sources stating the same fact
  - Accuracy of the sources
- We do not suppose we have any information on the facts truth
- Real world applications : query answering, source selection, data quality assessment on the web

Model

- Set of facts F = $\{f_1 \ldots f_n\}$
- A view is a partial mapping from F to {true, false}
- We work with a set of views V = $\{V_1 \ldots V_n\}$
- Our goal is to find the most likely real world (a total mapping) given the set of views.

Model : probabilistic model

- Hidden parameters :
  - $W : F \rightarrow \{True, False\}$
  - $\varepsilon : V \cup F \rightarrow [0,1]$, error factor
  - $\phi : V \cup F \rightarrow [0,1]$, ignorance factor
- Model :
  - $P(V_i(f_j) \text{ is undefined}) = \phi(V_i) * \phi(f_j)$
  - $P(V_i(f_j) = \neg W(f_j)) = (1- \phi(V_i) * \phi(f_j)) * \varepsilon(V_i) * \varepsilon(f_j)$
  - $P(V_i(f_j) = W(f_j)) = (1- \phi(V_i) * \phi(f_j)) * (1- \varepsilon(V_i) * \varepsilon(f_j))$
- More complex model : ignorance factor depends of the “belief” of the view
- Classical statistic methods (e.g. EM) are not directly applicable on this model because of non-linearity and high number of parameters

Algorithms : Base-lines

- Voting : more views pro than against

$$\Theta(W(f_j)) = \begin{cases} T \text{ if } \dfrac{|\{V_i, V_i(f_j) = T\}|}{|\{V_i(f_j) \text{ defined}\}|} \geq 0.5 \\ F \text{ otherwise} \end{cases}$$

- Counting : popular facts

$$\Theta(W(f_j)) = \begin{cases} T \text{ if } \dfrac{|\{V_i, V_i(f_j) = T\}|}{\max_f |\{V_i(f) = T\}|} \geq 0.5 \\ F \text{ otherwise} \end{cases}$$

- Page Rank : Counting in undirected graphs

Algorithms : fix-point intuition

1. Estimate the truth of the facts (eg. with voting)
2. Based on that estimate the error of the sources
3. Based on that refine the estimation for the facts
4. Based on that refine the estimation for the error of the sources
5. …

• Continue until a fix-point is reached (and cross your finger it converges)

Algorithms : 2-estimates

- Fix point algorithm on the following equations

$$\Theta(W(f_j)) = \frac{\sum_{V_i(f_j)=T} 1-\Theta(\varepsilon(V_i)) + \sum_{V_i(f_j)=F} \Theta(\varepsilon(V_i))}{|\{V_i, V_i(f_j)\ defined\}|}$$

$$\Theta(\varepsilon(V_i)) = \frac{\sum_{V_i(f_j)=T} 1-\Theta(W(f_j)) + \sum_{V_i(f_j)=F} \Theta(W(f_j))}{|\{f_j, V_i(f_j)\ defined\}|}$$

A fact is true
- if a view states it is true and make no error
- or if a view states it is false and make an error

A view make an error
- if it states a fact is true and the fact is false
- if it states a fact is false and the fact is true

- Instability $\Rightarrow$ tricky normalization

Algorithms : cosine

- Fix point algorithm on the following equations

$$\Theta(W(f_j)) = \frac{\sum_{V_i(f_j)=T} \Theta(\varepsilon(V_i))^3 - \sum_{V_i(f_j)=F} \Theta(\varepsilon(V_i))^3}{\sum_{V_i(f_j)\ undefined} \Theta(\varepsilon(V_i))^3}$$

$$\Theta(\varepsilon(V_i)) = \frac{\sum_{V_i(f_j)=T} \Theta(W(f_j)) - \sum_{V_i(f_j)=F} \Theta(W(f_j))}{\sqrt{|\{f_j, V_i(f_j)\ defined\}| \sum_{V_i(f_j)\ defined} \Theta(W(f_j))^2}}$$

- The truth of the fact is what the views state weighted by (the cube of) how error prone they are.

- The error of a view is the correlation between its statements on the facts and the predicted value for these facts.

Algorithms : 3-estimates

- Fix point algorithm on the following equations

$$\Theta(W(f_j)) = \frac{\sum_{V_i(f_j)=T} 1-\Theta(\varepsilon(V_i))\Theta(\varepsilon(f_j)) + \sum_{V_i(f_j)=F} \Theta(\varepsilon(V_i))\Theta(\varepsilon(f_j))}{|\{V_i, V_i(f_j)\ defined\}|}$$

$$\Theta(\varepsilon(V_i)) = \frac{\sum_{V_i(f_j)=T} \frac{1-\Theta(W(f_j))}{\Theta(\varepsilon(f_j))} + \sum_{V_i(f_j)=F} \frac{\Theta(W(f_j))}{\Theta(\varepsilon(f_j))}}{|\{f_j, V_i(f_j)\ defined\}|}$$

$$\Theta(\varepsilon(f_j)) = \frac{\sum_{V_i(f_j)=T} \frac{1-\Theta(W(f_j))}{\Theta(\varepsilon(V_i))} + \sum_{V_i(f_j)=F} \frac{\Theta(W(f_j))}{\Theta(\varepsilon(V_i))}}{|\{V_i, V_i(f_j)\ defined\}|}$$

- The difference with 2-estimate is that we take in account how hard a fact is, i.e. how likely the views are to make an error on the fact.
- More instability $\Rightarrow$ more tricky normalization

Functional dependencies

- What is FD for us?
  - One true value for a query among a set of values
- How to use it as a pre-filtering ?
  - When a view states true for a value, it states false for the other values
- How to use it as a post-filtering ?
  - Choose the best answer among the true ones (but keep hierarchy between “false” answers)

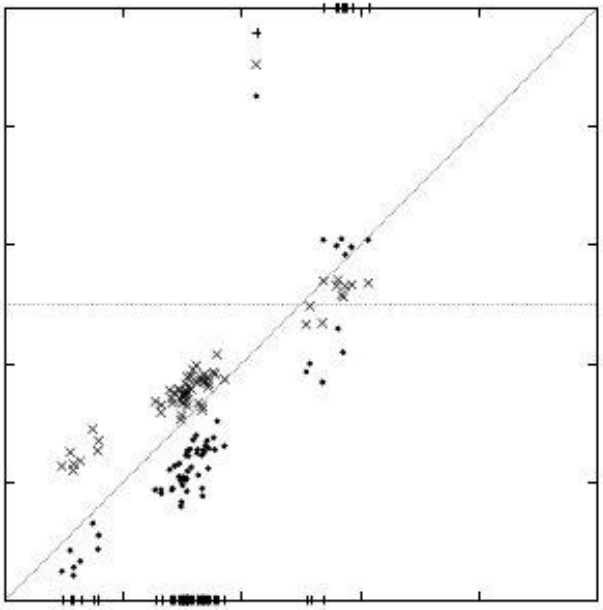
Experiments : what to measure?

- Quality of binary classification : percentage of error of prediction of the truth.
- Comparison between correctness of the fact (ie. real percentage of errors of the views on the fact) and confidence on the estimation of the truth of the fact
- Comparison between correctness of the view (ie. real percentage of errors of the view on the facts) and estimation of this value
- Precision-Recall if ordering fact by confidence on the fact that they are true
- Synthetic data-set generation using the full possibilities of our probabilistic data model

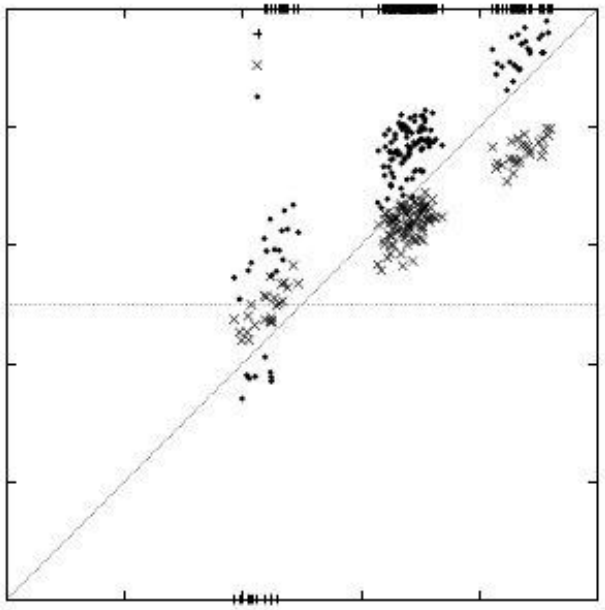
Experiments : synthetic data set

- Very good result of the binary separation for base-line methods but large improvement with the fix-point algorithms, specially 3-Estimates
- Different behaviors : 2-Estimates : binary separation, cosine : linear separation, 3-estimates : clear linear separation

+ 2-estimates
× Cosine
• 3-estimates



False facts



True facts

Experiments : real data set

- Three Real-world data-set :
  - General knowledge quiz (17 questions, 4 to 14 answers/question. 601 views)
  - Sixth-grade biology test (15 questions, true/false answers, 86 views) with semantic functional dependencies
  - Search Engine Queries (50 keyword queries, 13 search engines)
- Varying performance of the technique
  - Best if the views differ in quality – gives more weights to facts stated by “good” views
  - Best with many views and many facts
  - More difficult if the views are bad quality
  - More difficult if there are hidden correlations between facts

Conclusion

- Cool work, since unsupervised learning is somewhere magic
- Connected with data management, but closer to data mining
- Base-line techniques work reasonably
- Surprisingly, we can improve
  - Delicate and depends on data set
  - Hard because non-linear models with high number of parameters leads quickly to complexity and instability of algorithms
- Many interesting perspectives
  - More on FD and multi-answer (emails, phone numbers)
  - Specialized sites and domain expertise (NASA site good for astronomy)
  - Time –dependent answers (old phone number vs. recent ones)
  - Use of ontologies (answers such as IdF and Île-de-France)