CSEE @ UNIVERSITY OF MARYLAND BALTIMORE COUNTY (UMBC)

Using A Knowledge Base to Advice Machine Learning Algorithms

Ashwinkumar Ganesan

October 19, 2014

1 INTRODUCTION

Today, The Web is an amalgamation of different kinds of content from websites containing information on topics, to audio and videos and social networks to aid human communication. With an aim to improve searching of web content and automate services over the web, the semantic web structure was introduced [3]. This semantic web's structure, makes the Web accessible and interpretable to computer systems. Linked Open data provides semantics to web content. Resource Description Framework (RDF) was used to construct DBpedia [2] (which contains Wikipedia data [1]). DBpedia is an example knowledge base (KB) which can be used to improve machine learning algorithms by providing statistical methods with semantic information about the input data.

2 NAIVE BAYES

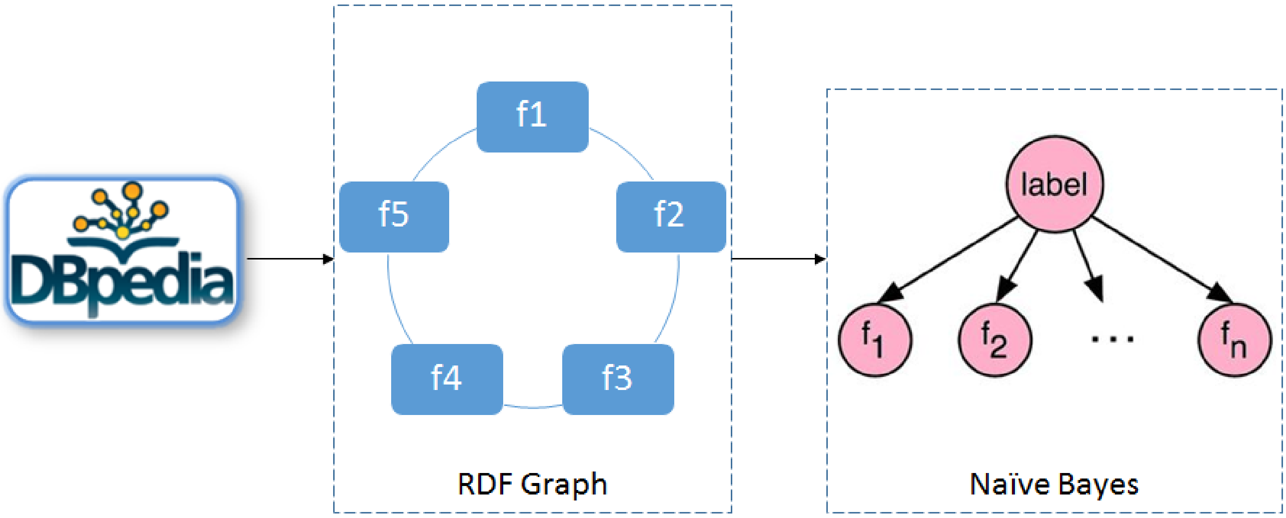

Figure 2.1: Sample workflow for *Naive Bayes*

Naive Bayes is a probabilistic classifier which applies the bayes theorem. *Naive Bayes* constructs a joint probability table of all the attributes, it is using for the purpose of classification. The method makes the assumption that there is conditional independence between variables. The workflow in the figure above [2.1] shows a sample workflow using the method and a knowledge base. Consider a data set which contains the variables *f1...fn*. A knowledge base (such as DBpedia) can be searched for attributes *f1..fn*, to get a connected graph showed how the attributes are linked and connected. This interlinking is used to inform *Naive Bayes* that variables *f1* and *f2* may not be conditionally independent. This can be used to create a joint variable. Also, the knowledge base graph provides other information such as the degree to which the two attributes (entities) are linked and how they linked. The attribute in the training set, can be combined with the data to give us a set of derived attributes which can be additionally used in the method.

3 Updating the knowledge base

As the knowledge base helps to improve the classifier, the test data can be used to update the links between entities in the knowledge base. The test data can be used to update missing values in the knowledge base of the attributes used in the statistical model. This raises important questions such as the data provenance as it may corrupt the knowledge base. The other option is to use Probabilistic Latent Semantic Analysis (PLSA) [4] to find correlated variables in the training / test data.

4 The Larger Picture

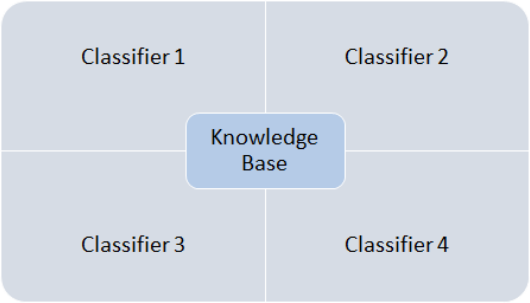

Figure 4.1: *The Mind Framework*

The above diagram [4.1] shows the larger framework where the knowledge base is used a passive entity by a multiple of classifiers. The framework tries to accomplish the following:

1. A method by which a classifier can use a knowledge base to understand the semantics of the data set and improve classification accuracy.

2. Use the test data set to improve the interlinking between entities in the knowledge base.

3. A knowledge base acts as a intermediate for classifiers to interact with each other.

References

[1] Wikipedia. Accessed: 2014-10-18.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. *Dbpedia: A nucleus for a web of open data.* Springer, 2007.

[3] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[4] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.