Text Classification

Special Topic Seminar

Presented by:

Mohammed Arshad Ansari

Project Guide:

Sharvari Govialkar

DEPARTMENT OF INFORMATION TECHNOLOGY
PILLAI INSTITUTE OF INFORMATION TECHNOLOGY
ENGINEERING, MEDIA STUDIES AND RESEARCH
NEW PANVEL - 410 206
UNIVERSITY OF MUMBAI
Academic Year 2014-15

Outline

Motivation

Following are the enumeration of causes that motivated this work.

1. Commercial need for fast and easy to implement classification mechanism
2. Unstructured data categorization
3. Reduction of manual time investment
4. Cost effectiveness

All the above mentioned needs and causes for motivation comes from the immediate requirement in the companies that heavily depend on the reselling.

Automatic document classification

1. ***Supervised document classification***; where some external mechanism (such as human feedback) provides information on the correct classification for documents,
2. ***Unsupervised document classification***; (also known as document clustering), where the classification must be done entirely without reference to external information, and
3. ***Semi-supervised document classification***: where parts of the documents are labeled by the external mechanism.

Existing techniques for classification

1. *Expectation maximization (EM)*
2. *Naive Bayes classifier*
3. *tf–idf*
4. *Latent semantic indexing*
5. *Support vector machines (SVM)*
6. *Artificial neural network*
7. *K-nearest neighbour algorithms*
8. *Decision trees such as ID3 or C4.5*
9. *Concept Mining*
10. *Multiple-instance learning*
11. *Natural language processing approaches*

Problems under consideration

1. Find similarity between products (at individual level) to group them together as same products and allow price differenciation.
2. Find similarity between products at a macroscopic level to group them together as belonging to same category of products automatically.
3. Find cross related products, to group them based on their function. E.g. Grouping similar pants and shirts together.

Outline

Literature Survey

Paper	Author	Approach	Result
A Machine Learning Approach for Automatic Text Categorization	Kurt Maly, Steven Zeil, Mohammed Zubair, Naveen Ratkal	SVM is used to classify the Defence Technical Information Center documents in to their appropriate category	SVM out perform other techniques such as bayes and KNN classifiers as shown in this work

Literature Survey

Paper	Author	Approach	Result
Automatic Text Classification: A Technical Review	Mita K. Dalal, Mukesh A. Zaveri	Multiple techniques are used for classification and evaluated in this work.	Stress must be given on feature selection, size and quality of training data to influence accuracy and correctness.

Literature Survey

Paper	Author	Approach	Result
Online News Text Classification Using Neural Network and SVM	Raghvan Gachli	Techniques such as Backpropagation Algorithm and SVM are used for text classification in this work.	Credence is given to the mixed (hybrid) approach for classification.

Literature Survey

Paper	Author	Approach	Result
Appyling Machine Learning to Product Categorization	Sushant Shankar and Irving Lin	Multiple techniques evaluated such as Naive Bayes, SVM and KNN.	Different data set yeild different results. Classification based on category set with ranking is given more suitablity in case of heirachy of categories

Literature Survey

Paper	Author	Approach	Result
Building semantic kernels for Text Classification using Wikipedia	Pu Wang and Carlotta Domeniconi	Bag of Words approach is improved by applying the knowledge from wikipedia to the sematic kernals that will be used by the BOW technique for a more informed classificati-	The overall bag of words based performance is enhanced due to more promity between synomyms, which are derived from the semantic kernals.

Literature Survey

Paper	Author	Approach	Result
GoldenBullet: Automated Classification of Product Data in E-commerce	Y. Ding, M. Korotkiy, B. Omelayenko, V. Kartseva, V. Zykov, M. Klien, E. Schulten and D. Fensel	A system called GoldenBullet is explained and evaluation for the purpose of text classification.	It uses a hybrid approach of combining data mining and machine learning techniques. The yeild is somewhere between 70% to 98%

Review Literature

1. Practicality of purpose
2. Classification versus clustering
3. Reusability versus Parallelization
4. Domain Knowledge

Outline

Other topics associated with text classification

1. Text Mining
2. Information retrival
3. NLP

Outline

Definition of problem statement

The general text categorization task can be formally defined as the task of approx- imating an unknown category assignment function F : D x C -> 0, 1, where D is the set of all possible documents and C is the set of predefined categories. The value of F(d, c) is 1 if the document d belongs to the category c and 0 otherwise. The approximating function M : D x C -> 0, 1 is called a classifier, and the task is to build a classifier that produces results as “close” as possible to the true category assignment function F.

Document Representation

The common classifiers and learning algorithms cannot directly process the text doc- uments in their original form. Therefore, during a preprocessing step, the documents are converted into a more manageable representation. Typically, the documents are represented by feature vectors. A feature is simply an entity without internal struc- ture – a dimension in the feature space. A document is represented as a vector in this space – a sequence of features and their weights.

1. Feature Selection
2. Dimensionality reduction

Feature Selection

1. Information Gain
   Entropy based information content quality mapping for finding interdependence of data.

$$IG(w) = \sum_{c \in C \bigcup \bar{C}} \sum_{f \in \{w, \bar{w}\}} P(f, c) \frac{P(c|f)}{P(c)}$$

2. CHI Square Method
   measures the maximal strength of dependence between the feature and the categories.

$$x^2_{max} = max_{c \in C} \frac{\| Tr \|.(P(f, c).P(\bar{f}, \bar{c}) - P(f, \bar{c}).P(\bar{f}, c))^2}{P(f).P(\bar{f}).P(c).P(\bar{c})}$$

Dimenionality Reduction by Feature Extraction

Dimenionality reduction relates to reducing the very large size of dimensions due to innumerable amount of possible words in a document.
Following are the methods with which we achieve the dimenionality reduction:

1. Synomym bag of words to reduce number of unique words/dimensions.
2. Term clustering for synonym grouping
3. Latent Symantic Indexing

Machine Learning based approaches for text classification

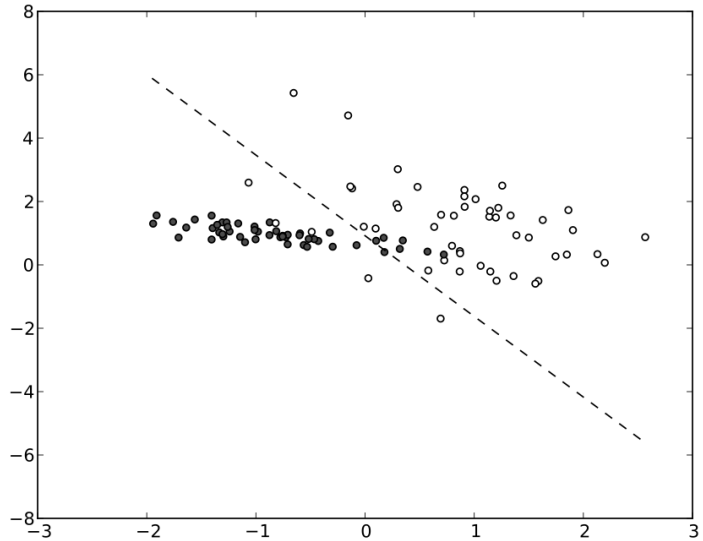
1. Categorization
    - 1.1 Naive Bayes Classifier
    - 1.2 Support vector machines
2. Clustering
    - 2.1 K Nearest Neighbor
    - 2.2 Expectation Maximization Algorithm

Naive Bayes Classifier

A

naive Bayes classifier assumes that the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 3ïn diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features.

Naive Bayes Convegence



Naive Bayes Classifier *contd.*

Simplicity of naive bayes classifier lies in the fact that its based on conditional property of bayes with the markov like process.

$$p(C|F_1 \cdots F_2)$$

$$p(C|F_1, \ldots, F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i|C)$$

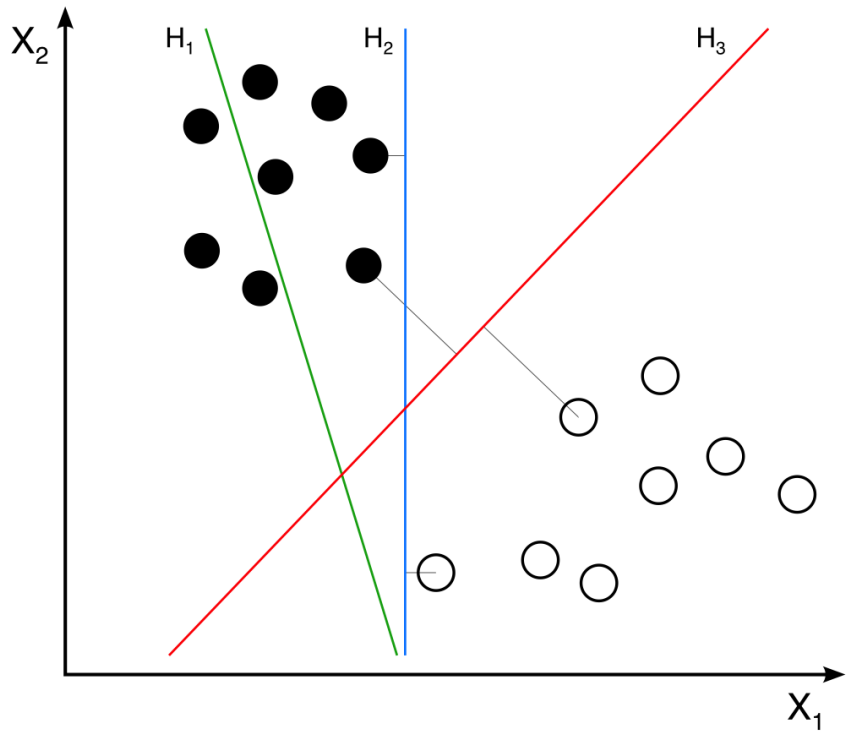
where the evidence $Z = p(F_1, \ldots, F_n)$ is a scaling factor dependent only on $F_1, \ldots, F_n$, that is, a constant if the values of the feature variables are known.

Support Vector Machine

A

support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

SVM Hyperplane

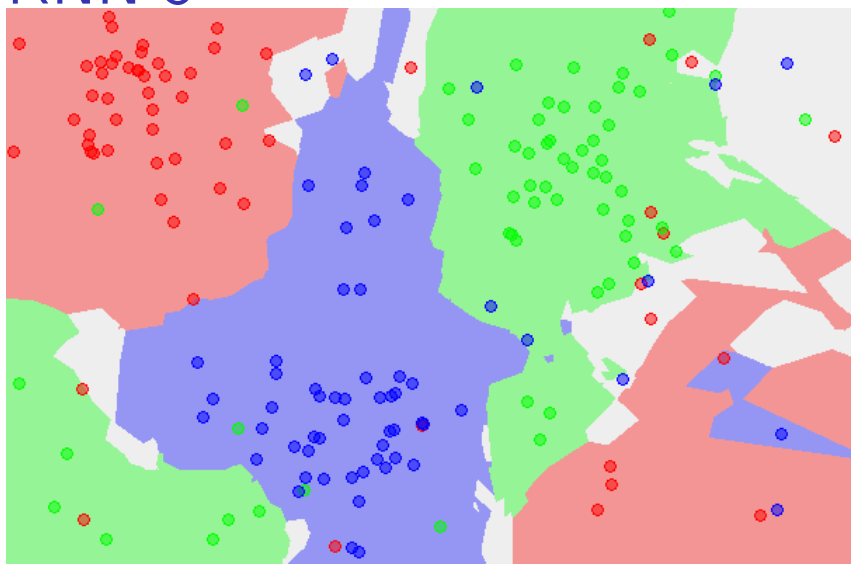


K Nearest Neighbor

A

KNN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

KNN 5



KNN *Contd.*

In k-NN regression, the k-NN algorithm is used for estimating continuous variables. One such algorithm uses a weighted average of the k nearest neighbors, weighted by the inverse of their distance. This algorithm works as follows:

- Compute the Euclidean or Mahalanobis distance from the query example to the labeled examples.
- Order the labeled examples by increasing distance.
- Find a heuristically optimal number k of nearest neighbors, based on RMSE. This is done using cross validation.
- Calculate an inverse distance weighted average with the k-nearest multivariate neighbors.

EM Algorithm

The EM algorithm is used to find the maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly.
Typically these models involve latent variables in addition to unknown parameters and known data observations. That is, either there are missing values among the data, or the model can be formulated more simply by assuming the existence of additional unobserved data points.
For example, a mixture model can be described more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component that each data point belongs to.

Hybrid approach

All the models can be used in combination with one another to achieve a common goal, while giving each a subtask.
Classfication and Clustering to the rescue.

- Need to use clustering when labels are not already known or present. But this can also be used to find synonyms amoungst the label.
- The label synomym thus obtained can be combined and the learning of Naive Bayes will be sufficient to identify new incoming documents.

Outline

General Applications

- **Spam Filtering** a process which tries to discern E-mail spam messages from legitimate emails
- **Email Routing** sending an email sent to a general address to a specific address or mailbox depending on topic
- **Language Identification** automatically determining the language of a text
- **Genre classification** automatically determining the genre of a text (also the objective of this work)

General Applications [*Contd.*]

- **Readability assessment** automatically determining the degree of readability of a text, either to find suitable materials for different age groups or reader types or as part of a larger text simplification system
- **Sentiment analysis** determining the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.
- **Article triage** selecting articles that are relevant for manual literature curation, for example as is being done as the first step to generate manually curated annotation databases in biology.

Ecommerce Application

1. **Product duplication across retailers**; There is a very obvious need to identify products being entered to see if they are duplicates. Especially for price comparison across retailers.
2. **Taxonomoy categorization**; Many products that belong to same category are put in the category they belong to, even if a similar to synonymous category exists. This causes multiple groups to be created by different word usage.
3. **Search enhancements**: Every time a product shows the signs of similarity in one way or another, it becomes easier to return search results based on this similarity, which in turn improves the overall experience.

Outline

Conclusion

Following were the observations that were made in the evaluation of all the literature that was considered for this work:

- **Size of data matters**: Small data set *HighBias/LowVariance* is most suitable for Naive bayes classifer since there is no overfitting. Large data set *LowBias/HighVariance* is most suitable for KNN or logistic regression.
- **Better feature dimensionality algorithm**: it is imperiative to use a feature dimensionality algorithm that reduces the number of words as much as possible. This gives best results for perforamce requirements and complexity reduction.

Conclusion [*Cont.*]

- **Hybrid Approach**: There are many aspects of each techniques that fit different problems in combinations. For example, clustering will help label the unknown faster, where as categorization will label individual documents faster.
- **Performance**: Performance requirements varies as per the problem domain. In some cases, traning the data set and then run against all the test data is done and in some cases, individual data is considered for classification at a time. This puts pressure on the type of algo chosen.

Outline

References

- Text Classification, Wikipedia,
- Automatic Text Classification, Internation Journal of Computer Applications (0975 - 8888)
- Online News Text Classification Using Neural Network and SVM, Raghvan Gachli, International Journal of Latest Scientific Research and Technology 1(2), July - 2014, pp. 122-128, ISSN: 2348-9464
- Applying Machine Learning to Product Categorization, Sushant Shankar and Irving Lin, Department of Computer Science, Standford University.
- Building Semantic Kernels for Text Classification using Wikipedia, Pu Wang and Carlotta Domeniconi, Department of Computer Science, George Mason University.

References [*Contd*.]

- GoldenBullet: Automated Classification of Product Data in E-Commerce, Y. Ding, M. Korotkiy, B. Omelayenko, V. Kartseva, V. Zykov, M. Klien, E. Schulten and D. Fensel., Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, NL.
- Text Minign, Wikipedia
- Information Retrieval, Wikipedia
- Natural Language Processing, Wikipedia
- Naive Bayes Algorithm, Wikipedia

References [*Contd*.]

- Support Vector Machine, Wikipedia
- K Nearest Neighbors, Wikipedia
- Expectation Maximization, Wikipedia
- Numerical example to understand Expectation-Maximization, StackExchange.com