Venciya.A, International Journal of Computer Science and Mobile Applications,
Vol.2 Issue. 12, December- 2014, pg. 87-96 ISSN: 2321-8363

Unsupervised Learning for Text Mining and Object Recognition in Images

Venciya.A
Student, Department Of Computer Science and Engineering, SRM University, Venciya.a@gmail.com

Abstract: ***In today's technology driven world where most information is stored digitally, being able to efficiently retrieve and sort the available information becomes highly crucial. Supervised Learning provides solution to some of these tasks, but they are highly dependent on human interaction for categorization of texts and other attributes in the initial stages. Unsupervised Learning methods on the other hand not only provide a viable solution to some of these tasks, but they also decrease the time taken to perform various tasks by removing the need for human interaction. In this paper we shall discuss an efficient way for text mining using feature extraction, feature selection, clustering and cluster evaluation. We shall then go on to discuss an efficient way of recognizing objects using Markov's Random Field.***

Keywords: Unsupervised Learning, Text Mining, Clustering, Object Recognition, Markov's Random Field

1. Introduction

With the advent of new hardware and communication technologies, there is an overwhelming amount of data that requires to be processed. The challenges that we face involve analysis, capture, search, retrieval, transfer, storage, and manipulation of data. While traditional methods are no longer viable to process such unorganized collection of data. Machine learning allows us to use probabilistic graphical models to determine the underlying structure of the data and then process data accordingly. In this paper, an overview of the various types of learning is provided. We also discuss an efficient way for proceeding with text mining along with the use of Markov's Random process for object recognition in images.

2. Machine Learning

Machine Learning deals with the construction and observation of algorithms that enables one to learn from a collection of data using predictions or decision rather than utilizing a fixed explicitly programmed sequential instruction set. Tom M. Mitchell 's widely accepted definition of machine learning is as follows "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". Thus a performance improvement based on past experience is what constitutes machine learning.
A Machine Learning System consists of the following components

- Goal: It is task specific and is defined as the intended result.
- Model: A model is a function, which maps perception to actions.
- Learning rules: Learning rules modify/alter the model parameters with new results and observations so that the performance with respect to the goals is optimal.
- Experience: A set of perception and the corresponding actions.

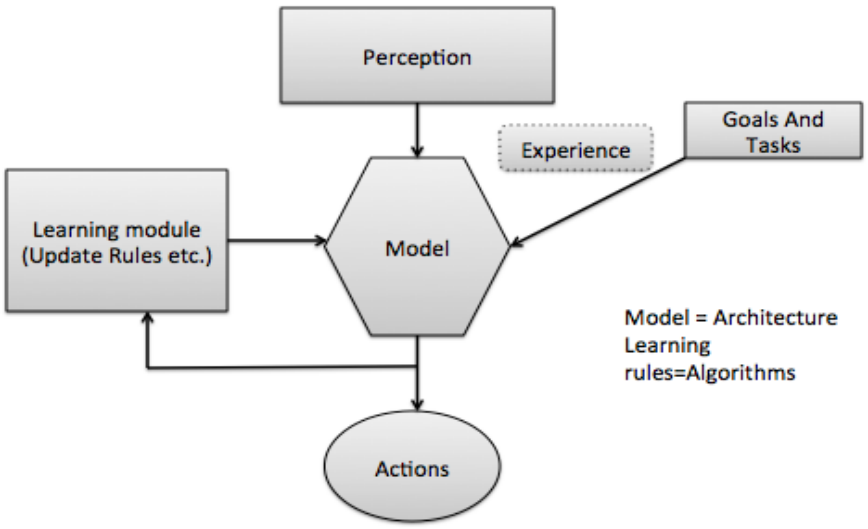

Figure 1: A schematic diagram of a generic learning system

3. Taxonomy of Learning Systems

3.1) Supervised learning:

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data is basically composed of training examples. And each example in turn consists of an input object and desired output value. The algorithm examines the training data and generates an inferred function, which is utilized for mapping new examples.

3.2) Reinforcement learning:

Reinforcement learning involves interaction with an environment. A reinforcement learning agent learns from the outcome of its actions, instead of being explicitly taught and it selects its actions based on its past experiences (exploitation) and also by new choices (exploration), which is basically a form of trial and error learning.

3.3) Unsupervised learning:

In unsupervised learning the learner discovers persistent patterns in the data consisting of a collection of perceptions. This is also called exploratory learning. Lets take a machine or any living thing which receives a sequence of inputs w1, w2, w3, w4 . . . where wt is the sensory input at a given time t. This input, which we will often call the data, could correspond to an image on the retina, the pixels in a camera, or a sound waveform. In unsupervised learning the machine receives the input and doesn't get any feedback from its environment. In a sense, unsupervised learning can be thought of as finding patterns in the data. This is accomplished using clustering and dimensionality reduction. Various basic models are used in unsupervised learning. Some of them are factor analysis, PCA, mixtures of Gaussians, ICA, hidden Markov models, state-space models and its variants or extensions. In this paper we discuss the application of Gaussian Markov Random Field to object recognition and an efficient way of using feature extraction, clustering, cluster evaluation in text mining.

4. Text Mining using Unsupervised Learning

In this paper an automatic text mining process that takes place in four stages is discussed. The four stages of text mining are

1) Feature Extraction

2) Feature Selection

3) Clustering

4) Cluster Evaluation

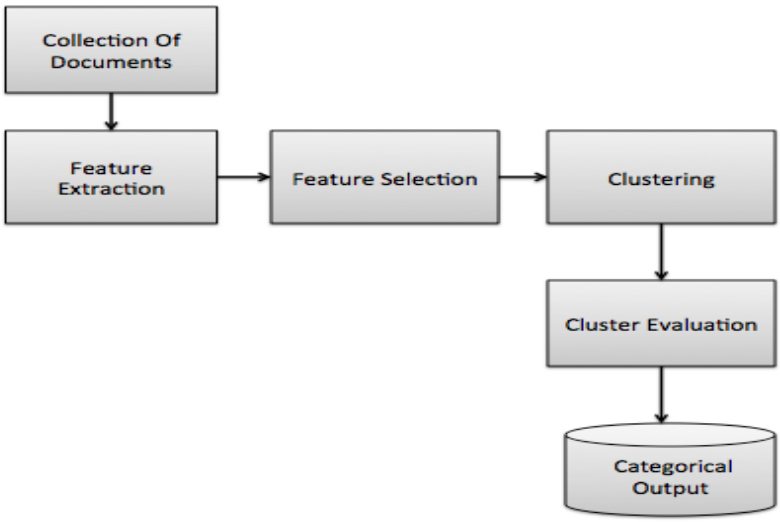

Figure 2: A schematic Diagram of the Automatic Text Mining Process.

4.1) Feature extraction:

Determining the important features to be extracted from the text plays a vital role in the success of text mining methods. As there is no meaningful way to combine and group text based on raw textual input, it is crucial to extract quantitative properties of the text that is most helpful in determining the class properties. We shall now discuss the unigram model and a recursive data-mining model along with document frequency, term strength and term contribution.

4.1.1) Document frequency (DF):

Document frequency takes into account the count of the documents in which a term occurs in a dataset. It is simple and scales well with high efficiency and hence is widely used in feature extraction. It is widely employed in text mining.

4.1.2) Term strength (TS):

Term strength is based on the assumption that if a term occurs in the first half of a document it would also occur in the second half of the related documents.

4.1.3) ***Unigram:***

In unigram model each word or stem is taken into account individually. Each word is considered to be atomic. Here feature extraction is wholly dependent on word frequency. Lets take an example where we need to determine the type of journal where a particular article was published. As we know there are common words and there are subject specific words associated with individual journals. Therefore a classification can be made based on presence or absence of certain words. This model can be extended and is known as bigram model, which focuses on occurrence of word pairs, instead of focusing on each single word. For example take the word machine learning; we see that the unigram approach considers 'machine' and 'learning' as separate entities whereas 'machine learning' is considered as a single entity in bigram approach. This increased expressive nature of bigram comes at a significant cost as there are many more pairs of words that will be observed in the text than just words themselves, thus resource computation tends to be more. Unigram and bigram can be extended to an n gram approach where n corresponds to the number of words that are to be grouped. But the cost and complexity of n gram increases with growing value of n.

4.1.4) RDM-Recursive Datagram Approach:

Recursive Data Mining records not only the words and related frequencies but determines the patterns and structures present in text. In this paper we present a recursive data mining technique where the patterns are not pre-determined but are discovered when the text is processed automatically. RDM receives a series of tokens and entities. The meaning underlying the tokens is irrelevant and their method of generation needn't be taken into account. Tokens can be generated using individual symbols, syllables, words etc. One generated they are fed to RDM to determine tokens that occur frequently. These sequences are individually assigned new token identifiers. All instances of the sequence are replaced with the corresponding new tokens and this is repeated till there are no frequent sequences. In this way sequences are built out of smaller sequences thus following a hierarchical model of text and the tokens are exported. Minor discrepancies can also be included allowing for distinct but originally similar sequences to be considered the same. Thus the overall number of sequences is reduced. And also the sequences that remain have a better probability of crossing the frequency threshold for it to be included in other potential sequences.

4.1.5) ***Term contribution:***

In this paper a new feature selection method called "Term Contribution" that takes the term weight into account is introduced. The result of any clustering method is dependent on the similarity among the documents. In order to calculate the similarity the procedure is as follows

$$Sim(dj,di)=\sum f(t,dj) \times f(t,di)$$

[f (t,m) represents the tf * jdf weight of terms in document m].

The term contribution is thus given by

$$TC(t)=\sum f(t,dj) \times f(t,di)$$

If weights are equal then TC (t) = DF (t)(DF (t) −1), thus Document frequency is a special case of Term contribution.

4.2) Feature Selection:

It is often seen that features extracted earlier might have to be filtered based on the context (i.e.) based on the specific text mining application. In this stage features that don't provide new information is discarded. Depending on the need of the application we shall now look at three feature selection methods.

4.2.1) Information gain and gain ratio:

Information gain is defined as IG(Class, attribute)=H(Class)-H(Class/Attribute) where H is the entropy and is given by $H(X)=-\sum Pr(X) \ln Pr(X)$

As soon as IG is calculated, a threshold can be determined so that every point exceeding the threshold is selected to be a part to the cluster, and everything else is ignored. If the entropy value is high the gain ratio parameter is used for clustering .The gain ratio is given by

Gain_Ratio (Class, Attribute) = Information_Gain (Class, Attribute) /H(Attribute) = H(Class) − H(Class| Attribute)/H(Attribute) .

4.2.2) *Principle component analysis:*

Principle Component Analysis is an unsupervised method for feature selection as no class information is required. PCA transforms the features that were selected into an entirely different set of attributes. An additional property of PCA is finding the vector that corresponds to the largest variance in the data. This vector may correspond to a combination of features. This vector in turn serves as a new attribute. Each data instance has a value associated with this new attribute determined by its projection onto the vector. The next attribute resulting from PCA is created by selecting a vector orthogonal to the previously chosen one that corresponds to the largest remaining variance, and then each data instance is projected onto this new vector. This process is repeated iteratively by selecting a new vector orthogonal to all vectors selected previously and projecting the data instances down to determine their value, resulting in a set of new features and their related values.

4.3) Clustering:

Clustering involves the division of a set of objects into clusters (parts of the set) so that objects in the same cluster are similar to each other, and objects in different clusters are dissimilar. Thus they are grouped into the best combination of objects. The cluster assignment that results map any data point to a cluster of those points most similar. A cluster is assigned points such that if a particular point is assigned to it, that same point cannot be assigned to another cluster. The points of a cluster must be similar to each other.

4.3.1) Expectation maximization

Expectation maximization by varying a probability distribution and its parameters attempts to determine the proper allocation of points to clusters. Every point has a probability distribution that lists the probability values of points being in one of the specified clusters. The points are then distributed according to maximum probability. Before distribution of the probabilities we need to 1) Keep parameters static and finding the best distribution and 2) optimize distribution with respect to parameters.

4.3.2) K-means

K means is an iterative clustering method that takes place as follows

1) Select n arbitrary points, which correspond to the centre of potential clusters.

2) Each data point is then assigned to the cluster of the centroid it is closest to, which creates n clusters.

3) The centre of each of those n clusters is then computed and these centres are considered the new centroids.

The above steps are iterated till n clusters are returned.

4.4) Cluster evaluation:

A cluster evaluation metric is used to determine the suitable number of clusters that are required for a given text mining problem. It is likely that many different clustering are possible. It allows for ranking of the clusters and selection of the cluster that is most accurately represented. The goal of cluster evaluation is to provide the single best clustering corresponding to the actual clustering .In this paper we consider two evaluation criteria. They are

4.4.1) Scatter separability:

Scatter separability is based on the assumption that a good clustering takes place only when all the points on a particular cluster are near each other while different clusters on the other hand should be far apart. Thus each point in a cluster must be similar while being radically different from points in other clusters. In order to measure scatter separability we first determine the intra cluster scatter matrix that represents the scatter points in a cluster and the inter cluster scatter matrix that is used as a representation of scattering among different clusters. The inverse of the intra-cluster scatter matrix is combined with the inter-cluster scatter matrix in order to compute scatter separability.

4.4.2) Silhouette coefficient:

The steps involved in calculating Silhouette Coefficient is as follows

1) The average distance between that point and all points in its own cluster is calculated .Let this be represented as i.

2) Find a point in another cluster that is closest to the point in a given cluster. Let this distance be j.

The silhouette coefficient is then defined as the Silhouette Coefficient =j – i

Thus predictions would be accurate if i is close to zero.

4.5) Pseudo code for experimental procedure:

PRE-REQUISITE: A collection of Documents.

1) For page1_size in [5,7,9] do

2) Data1_set ← Create Data1_set (page1_size)

3) for Feature1_Extractor in [Unigram, RDM] do

4) features all ← Feature1_Extractor(data_set)

5) for Feature1_Selector in [Information Gain, Gain Ratio, PCA] do

6) features selected ← Feature1_Selector(features all)

7) for Cluster in [EM_KMeans] do

8) clustering ← Cluster(features selected)

9) accuracy ← Compute_Accura(clustering, data_set)

10){

11) Record accuracy as a function of page1_size, Feature1_Extractor, Feature1_Selector, and Cluster

12) }

14) eval ← Cluster1_Eval(clustering)

15) {

16) Record eval as a function of pages1_ize, Feature1_Extractor, Feature1_Selector, Cluster, and Cluster1_Eval

17)}

18) end for

19) end for

20) end for

21) end for

22) end for

23) End

5. Markov's Random Field

Markov random field also known as the Markov network or undirected graphical model is a set of random variables displaying a Markov property described by an undirected graph.

A Markov Random Field is a graph G = (V, E).
• V = {1, 2,3,4…., N} is the set of nodes, each of which is associated with a random variable, u_i, for i = 1...N.
• The neighborhood of node j, denoted N j , is the set of nodes to which j is adjacent; i.e., i ∈ N j if and only if (j,i) ∈ E.

• The Markov Random field satisfies p(uj | {ui}j∈V\j) = p(uj | {ui}i∈Ni)
Nj is often called the Markov blanket of node j

6. Gaussian Markov Random Field

A multivariate normal distribution forms a Markov random field with respect to a graph $G = (V, E)$ if the missing edges correspond to zeros on the inverse covariance matrix.

7. Use of Markov Random Fields for Object Recognition:

In-order to recognize objects in an image or video sequence, we make use of a special case of the Markov random fields: the Gaussian Markov random fields where a given pixel in a given color channel is linearly dependent on its 26-connected immediate neighbors in both geometric and the standard RGB color space also known as Warp color space. The 3 × 26 model parameters are coefficients that dictate the linear relationship between a pixel and all its neighbors at the same time representing the colored texture.
The texture also known as conditional probability density function is defined as a zero mean 3D Gaussian function, whose covariance noise matrix is computed from the pixels in the region and the parameters. Thus the input value would be zero if a pixel and surrounding pixels were in accord with the texture parameters.
This proceeds in three phases
1) Region splitting
2) Conservative clustering
3) Stepwise optimal clustering

7.1) Regional splitting phase:

In this stage, the image is divided into smaller and smaller square regions until a uniform texture is obtained in a given region. The process must take place in such a way such that already uniform regions must not be segmented. The uniformity test consists of computing the mean color error of all pixels in the region, and comparing it with the mean color error of each of the four possible sub-square-regions. In order to determine if the region is split or not we determine the covariance. If the covariance of a region is below a certain threshold and if the difference is also below some threshold then the region is not split.

7.2) Conservative clustering:

In the conservative clustering step the adjacent regions that are similar are merged locally. This phase is used to reduce the workload for the next phase. To determine the similarity between two adjacent regions candidate for merging: the color mean differences need to be lower than a given threshold, with the covariance matrix being small and if the regions are not too small for pseudo-likelihood estimation a lower threshold must be set. All regions are processed till equilibrium is obtained. A stopping criterion is then used to finally stop the merging process that takes places in a sequence of iterative steps.

7.3) Stepwise optimal clustering:

Since our focus lies on the parameters of the texture, the PDF associated with each pixel is the likelihood of the parameters. Multiplying PDF values of all pixels in a given region provides a pseudo-likelihood. This is clearly is not a true likelihood as neighboring pixels are dependent of each other, but using this value for a maximum likelihood approach proved to give good results for parameter estimation.

For merging two potential adjacent regions, we continue by comparing the pseudo-likelihood of the whole image when the two candidate regions are left separate, and when they are merged. Subsequent adjacent regions to a region are likewise tested, and the pair with the maximum pseudo-likelihood wins, and are merged if the difference in global pseudo-likelihood, which will increase as merging progresses, is less than a given Stopping criterion.

Similarly, the other regions are processed until the stopping criterion is reached. If a region of the pair is found to be too small for pseudo-likelihood estimation, color mean difference tricks are used, but without the use of thresholds. In case of absence of better pairs they will be merged. Given its Markovian nature it can be easily implemented on a parallel architecture to obtain an improved performance.

8. Acknowledgement

I would like to thank SRM University, Kattankulathur and the CSE Department for their support.

9. Conclusion

Thus an efficient method for text mining based on feature extraction, selection, clustering and cluster evaluation is presented in this paper along with the use of Markovian Random field, more specifically Gaussian Markovian Random fields for object recognition in images.

References:

[1] Mitchell, T. (1997). Machine Learning, McGraw Hill, ISBN 0-07-042807-7, p.2.

[2] Youngjoong Ko and Jungyun Seo, Automatic Text Categorization by Unsupervised Learning, Pages 453-459 Volume 1 ISBN:1-55860-717-X.

[3] J. MacQueen. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1(281-297): 14, 1967.

[4] T. Moon. The expectation-maximization algorithm. Signal Processing Magazine, IEEE, 13(6): 47–60, 1996.

[5] K. Fukunaga. Statistical Pattern Recognition (2nd Ed). Academic Press, 1990.

[6] L. Kaufman and P. Rousseeuw. Finding groups in data. An introduction to cluster analysis. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, New York: Wiley, 1990, 1990.

[7] M. Caillet, J. Pessiot, M. Amini, and P. Gallinari. Unsupervised learning with term clustering for thematic segmentation of texts. Proceedings of RIAO, pages 648–660, 2004.

[8] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying," IEEE Trans. Pattern Anal. Mach. Intell. vol. 24, no. 8, pp. 1026–1038, 2002.

[9] M. Weber, M. Welling, and P. Perona, "Unsupervised Learning of Models for Recognition," Proc. 6th European Conference Computer Vision
(ECCV), pp. 18–32, June 2000.

[10] R. Fergus, P. Perona, and A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning," Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 264–271, June 2003.

[11] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) *Foundations of Machine Learning*, The MIT Press ISBN 9780262018258

[12] Rue, Håvard; Held, Leonhard (2005). *Gaussian Markov random fields: theory and applications*. CRC Press. ISBN 1-58488-432-0.

[13] D. K. Panjwani and G. Healey, "Unsupervised Segmentation of Textured Color Images," IEEE Trans. Pattern Anal. Mach. Intell., vol. 17, no. 10, pp. 939–954, 1995.

[14] D. Maynard, V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, and Y. Wilks. Architectural elements of language engineering robustness. Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data, 2002.

[15] G. Neumann, R. Backofen, J. Baur, M. Becker, and C. Braun. An information extraction core system for real world german text processing. In Proceedings of ANLP-97, pages 208–215, Washington, USA, 1997

[16] M. Reinberger and P. Spyns. Unsupervised text mining for the learning of dogma-inspired ontologies. Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press. 2005.

[17] Ying Liu, Dengsheng Zhang, Guojun Lu, M. Wei-Ying, "A survey of content-based image retrieval with high-level semantics". Patern Recognition, Volume 40 , Issue 1, p. 262-282 ,January 2007

[18] Z. Ghahramani. Unsupervised Learning. Advanced Lectures on Machine Learning, 3176:72–112, 2004.

[19] I. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques (2nd Ed). Morgan Kaufmann, San Francisco, 2005.

[20] L. Luchese and S.K. Mitra, "Color Image Segmentation: A State-of-the- Art Survey", Procedings of the Indian National Science Academy, 67(2): 207–21,March 2001.

