

A Survey of Machine Learning and Computer Vision with SVM

Travis Gneiting
travis801@hotmail.com
(801)891-3461
Chuck Anderson

Colorado State University
CS 545- DL – Machine Learning

December 16, 2009

Submitted as Assignment 8 for CS545 DL – Machine Learning

Keywords: Computer Vision, Image Recognition, Machine Learning, SVM

Abstract

Using machine learning to analyze and categorize photographs is a popular growing field of study. I will compare three papers that look at different methods for using and accomplishing image recognition with machine learning techniques. The three papers consist of techniques for image and text machine learning. Matching Words and Pictures [1] focuses on auto-annotation and region naming for images. Harvesting Image Databases from the Web [2] looks at an approach for capturing large datasets of images from web searches. Training Support Vector Machines: An Application to Face Detection [3] investigates the use of Support Vector Machines (SVMs) for the use of facial recognition.

1.0 Introduction

While enrolled in a CS545 Machine Learning at Colorado State, I studied many machine learning algorithms and approaches for machine learning. I found most fascinating the use of machine learning techniques to classify and analyze images. Additionally, I was interested in learning more about computer vision as a field for machine learning. I selected the three papers based on their relationship to computer vision and machine learning.

Computer vision seems to be a popular topic for machine learning; I think this is because of the broad opportunities that have yet to be researched in this area. The hand written zip code

recognition assignment was very interesting to me, and I wanted to learn more about computer vision with images.

In section 2.1, I review Matching Words and Pictures [1]. This paper explores the use of text and images to improve the searching capability of very large collections of images. The images are matched based on features that are shown in the images and categorized as different regions. This service could provide automated image annotation, image browsing support, and auto-illustration to provide suggestions for images to illustrate text.

In section 2.2, I evaluate the paper Harvesting Image Database from the Web [2] which investigates the use of text and images searches on the web using Support Vector Machines (SVM) and visual classifier to learn and improve rankings of images in internet searches. The ideas provided seek to harvest a large data set of images and text from the web to train and classify with. The goal was to produce highly precise result set using computer vision as well as text, meta-data and other visual information to provide the best rankings in a large database of images.

Section 2.3, focuses on the paper Training Support Vector Machines: An Application to Face Detection [3]. This paper uses SVM to aid in computer vision to help solve problems with facial recognition and detection. The problem focus was to train the classifiers to discriminate between face and non-face patterns in a data set of images. The solution is accomplished with the use of SVM to train the system. This paper was published over a decade ago, I wanted to review what might have changed in the use of SVM for computer vision applications in contrast to the other papers reviewed.

I will review the motivation for these reports, what approaches they took toward solutions, and the final results of each papers study.

2.0 Computer Vision (Summary of Papers)

2.1 Matching Words and Pictures

Words and images have a relationship when trying to locate an image based on a description or explanation of a picture. This paper takes an approach to use auto-annotation and image region naming to add additional meta data to aid in classifying and locating images. The joint distribution of different regions in an image includes areas for learning, and provides a correspondence between the regions of an image and matching words. A multi-modal and correspondence extension to Hofmann's hierarchical clustering/aspect model is used for annotating images of real scenes. A big problem with matching words and pictures is the performance; the time involved for large datasets is very expensive.

There are many practical applications that have motivated this work. For example, large databases of images that are hard to locate certain images based on filenames or directories. Automated image annotators were used in the application of processing the image data. In the past annotations of images in newspapers and journals were often difficult and hard to understand. A helpful procedure would be to automatically annotate these types of images. Museums often provide images of collections on the web. When a user wishes to search for these images it is often hard to find a group of the images together. Providing a collection of similar images groups with similarly annotated text would provide a valuable service for these collections. There are large online image commercial collections, which often do not provide enough information for users to locate the image they are seeking. A tool could be used to automatically suggest images that would be similar to text or copy that is provided as a query.

Before the images are processed, there was an amount of preprocessing that was completed to segment normalized cuts of the images. The segmenter typically divides an image into eight large regions to be computed by each region individually. Some of the analysis was done on the size of the image and what was covered by a region. Additionally, position, color, texture and shape were reviewed and recorded for the image.

Once it has been determined that an image is clear enough, all clusters were taken into consideration and a document was modeled over the sum of the clusters. The weights were then allocated based on the probability that a given portion of the image was in a cluster. These observations had an association with the document now providing text with regions of an image.

A model used to apply generated words was used. This model is referred to as Latent Dirichlet allocation (LDA) not to be confused with Linear Discriminate Analysis that was covered previously in our course work. The model was used to extend multi-modal data (MoM-LDA) that assumed each image and assigned words were correctly generated by the same process. Given an image and the MoM-LDA, a computation is then used to approximate the distribution over words given to an image. MoM-LDA derives its predictive abilities from higher levels components that weren't directly covered in this paper.

To further use the correspondence models, the authors of the paper wanted to build models that could predict words for specific regions. They used a hierarchical aspect models and MoM-LDA to produce the correspondence information.

There were many attempts to evaluate the models of different kinds of data. Measurements of well represented data sets were used in addition to data that was not well represented. The quality of the selected word predictions were estimated by the divergence between the computed predictive distributions. This was taken from the model and a dictionary to calculate the error for the given model.

Measurement for the models predicted words often containing various levels of errors. For example, “cat” may be substituted for “tiger” is a less offensive error than a substitution of “car” for “tiger”.

Training the information was shown to learn fairly quickly. Thirty iterations were showing the similar results as hundreds of iterations. Additionally, the paper indicated that over fitting was not a big problem. There was some human interaction to partition the images into positive and negative training classifiers based on visual shapes. This ensured that the classifiers were receiving good training data.

Overall, the paper provided a good sense of understanding of the problem and solution for Matching Words to Pictures. The results from smaller data sets only including a relatively simple classification problem returned very good results. The tagging of images with nouns was very well taken. However, with an increasingly large dictionary, the results began to degrade. An increase in supervised learning may have assisted in improved results, but the author did not discuss this.

Future work by the author will work with large dictionaries to analyze the correctness of the system.

2.2 Harvesting Image Database From the Web

The idea behind harvesting image database from the web is to automatically create a set of images for a specific class. Using search engines are often times hard to get a large data set of a given object being queried. The paper gives the example that a search for “shark” returned only 39% of images with relevancy to the query. The motivation of the paper is to provide a multi-modal for text, meta data, and visual features within the images to produce a large data collection of images from the web. The task is described as removing irrelevant images and re-ranking the remainder of images for relevancy. The images are ranked using a Bayes posterior estimator that was trained on any text that had to do with the image. This included image name, title tags or additional meta data. After the images were ranked based on the text using Bayes algorithm, the top ranked images were used as training data. A SVM visual classifier was also used for improved ranking. Similarly to the paper listed above, Latent Dirichlet Allocation (LDA) was used on the text for images downloaded for testing and training from the web.

A major challenge was the collection of a large data set of images, specifically from the web due to download restrictions from search engines and websites. Another large hurdle was how to combine text, meta-data and visual information to provide the best image ranking and re-ranking system. The main source of images for the data set was collected using web searches, Google searches and Google Image searches. These search results were used to collect both images and text information about the image. These images were then divided into various categories for

processing. It is important to point out that because of the difficulty involved in text associated with images, it is hard to know exactly how relevant the data set is based on an image and text association. An example was given that the search “shark” would also return team names (“San Jose Sharks”).

The main objective was to create a database of categorized images. Another challenge involved with the objective is the observation of abstract images and how to compare natural images and remove abstract images.

The main learning filter was trained using SVM. This was very accurate on the hand drawn images and symbols that were used. This was used to further classify or remove images from the data set to provide a cleaner collection of classified images.

Additional ranking of the images were done based on the text features of the images. Some of the highlighted features were the context of the website the image was extracted from, the file name, file directory, image alt tag, image title, and website title. These were all used to provide additional precision to the features of the image. Image ranking from these additional features was then applied to a Bayesian posterior estimation.

The final text based results were ranked using a text based posterior for ordering. When ranking results of textual and visual images, the classification performance was not affected much by noise of the training data used by SVM. It was stated that “SVM based classifier is shown to be very noise insensitive and well suited for this task”. Two other classification methods were reviewed for this solution, pLSA and a feature selection based approach. SVM showed the best performance and was selected for use as the classifier.

The final algorithm for image search database was compared against other approaches. Overall the precision from SVM offered a better solution. It also increased the ranking results, with the included manual intervention. The filter was successful for building a small sample database.

2.3 Training Support Vector Machines: An Application to Face Detection

SVM provides methods for training polynomial, neural networks, or Radial Basis Functions classifiers. There are many issues involved in solving for intense quadratic equations. This optimization problem is reviewed in this paper along with the ability for the memory to grow with the number of points in a data set. The presented algorithm for a solution is SVM to train the data set. The experiment for handling large data set is tested on a facial detection problem with a very large set of data points.

Facial detection problems involve large data sets. This causes problems and complexity when representing this data. One solution to improve performance that was used in this paper was to transfer the image to grayscale values.

The objective this paper focuses on the use of SVM and the use of a pattern classification algorithm. The paper praises SVM for its positive results when used to train polynomial, neural networks and Radial Basis Function classifiers.

Most techniques for training focus on minimizing error or empirical risk. SVM focus on structural risk minimization. Structural risk minimization spotlights the minimized upper bound of the generalization error. The use of SVM is similar to solving with linear constrained quadratic programming. The paper points out that this issue is caused by a data set become too large. The paper focuses on minimizing the quadratic programming by a decomposition algorithm. SVM is still used as the main algorithm for the face detection system.

The problem of training a large data set can be difficult. The paper provides a decomposition algorithm to divide the data in samples. An interesting solution was provided in the paper for working with large datasets. To decompose the large data set problem, they solved iteratively by keeping the system fixed at zero level and look for data points that were non support vectors, so the final optimizing was done on a reduced set of variables.

For improvements and the decomposition algorithm, the objective of the algorithm is to converge to the global optimal solution. The decomposition can be improved by adding points that violate the optimality conditions in the linearly separable margins.

The application of SVM to face detection was designed to be used for detecting vertically oriented and unoccluded frontal views of humans in grayscale. The paper selected the use of SVM on facial detection because of the challenge and potential for success with SVM. There are many potential issues with facial detection because of the many outside and changing variables.

Before this paper, there has been previous work done with reflect systems that have shown very high positive rates which encouraged the work of the paper. There was additional preprocessing for the facial recognition, to match size with the use of masking, illumination gradient correction, and histogram equalization to compensate for different camera response curves and lighting. Once the image has been selected misclassifications are stored as negative training examples. They refer to this as a bootstrapping step as a training phase. The training phase of the negative examples seemed to improve the definition overall, so the additional negative examples were found to be helpful.

The new ideas presented in this paper using SVM on large data sets seemed very new and advanced for the time this paper was written. Specifically because the difficulty with processing power being limited to the researchers at this time. SVM's have a solid mathematical foundation and proved to be a positive tool for analyzing the data points in the given facial detection problem. The paper proved that SVM handled high dimensional input vectors very well.

3.0 Conclusion and Observations

The main concept that stood out to me in these three reviewed papers was the use and praise of SVM. I specifically choose the last paper that is over a decade old to see the use and comparison of SVM's from that time until the use of SVM's today. The general techniques for classifying data and performing computer vision techniques on various data sets in each of the papers seemed to be very similar. The first and second paper reviewed did not cover as in-depth the algorithms and mathematical process of the classification as the third paper did. However, they each faced the similar issue of working with large scale data sets.

The uses of SVM were effective in these applications because of the value added when trying to classify data.

In papers one and two, the Latent Dirichlet Allocation (LDA) was used which has not been mentioned before. The purpose was to generate a model with the groups. I would be interested to read more topics of Latent Dirichlet Allocation and generative models. The act of modeling random observations of data in relation to the joint probability distribution seemed to have a lot of relative information to the machine learning techniques covered in our class. After further research, I noticed that LDA was not developed until 2002 by David Blei, Andrew Ng, and Michael Jordan [5]. This may have been why it was not mentioned in the third paper. Additionally, Matching Words and Pictures was authored by two of the founders of LDA which may explain the use and expansion of MoM-LDA.

The three reports were very supportive of each other. The similarities with constraints of large data sets and image data sets seem to always be an issue in training and learning algorithms. The approach for decomposing the large dataset problem used in the paper on training SVM could have been helpful for the other papers that also were faced with large data sets.

There was some difficulty understanding the complexities in the third paper. With more time I felt I could have better understood the deep explanations of SVM used in computer vision. The first and second paper did not offer a lot of in-depth explanations for the algorithms used and the calculations performed. I assume this may be to the wide acceptance of SVM and the familiarity of it within the machine learning community.

The three reports that were reviewed for this survey provided valuable information and ideas in the area of computer vision and image recognition. I plan to continue my studies on computer vision with images in the future. I learned a vast amount about the procurement and processing of large data sets, and the preprocessing techniques used to speed up training and learning techniques. I was also able to build on my understanding of SVM from the brief coverage from our class lectures.

References

- [1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching Words and Pictures. *J. Machine Learning Research*, 3:1107–1135, Feb 2003.
- [2] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting Image Databases from the Web. In *ICCV*, 2007.
- [3] E. Osuna, R. Freund, and F. Girosi. Training Support Vector Machines: An Application to Face Detection. In *Proceedings of CVPR'97*, pages 130–136, New York, NY, 1997. IEEE.
- [4] http://en.wikipedia.org/wiki/Computer_vision
- [5] http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
- [6] C. Anderson, CS545: Distributed Computing Using Snowfall and Support Vector Machines <http://www.cs.colostate.edu/~anderson/cs545/Lectures/week15day1/week15day1Handouts-2x2.pdf>