Document Classification with Support Vector Machines

Konstantin Mertsalov
Principal Scientist, Machine and Computational Learning
Rational Retention, LLC
kmertsalov@rationalretention.com

Michael McCreary
Chief Operating Officer
Rational Retention, LLC
mmccreary@rationalretention.com

January 2009

Introduction

Document classification is the task of grouping documents into categories based upon their content - never before has it been as important as it is today. The exponential growth of unstructured data combined with a marked increase in litigation, security and privacy rules have left organizations utterly unable to cope with the conflicting demands of the business, lawyers and regulators. The net is escalating costs and risks, with no end in sight. Without tools to facilitate automated, content based classification, organizations have little hope of catching up, let alone getting ahead of the problem. Technology has created the problem, and technology will be needed to address it.

Manual classification is out of the question due to the volume of data, while naïve automatic approaches such as predefined search terms have performed poorly due to the complexity of human language. Many advanced approaches have been proposed to solve this problem, however over the last several years Support Vector Machines (SVM) classification has come to the forefront. SVM's deep academic roots, accuracy, computational scalability, language independence and ease of implementation make it ideally suited to tackling the document classification challenges faced by today's large organizations.

SVM

SVM is a group of learning algorithms primarily used for classification tasks on complicated data such as image classification and protein structure analysis. SVM is used in a countless fields in science and industry, including Bio-technology, Medicine, Chemistry and Computer Science. It has also turned out to be ideally suited for categorization of large text repositories such as those housed in virtually all large, modern organizations.

Introduced in 1992, SVM quickly became regarded as the state-of-the-art method for classification of complex, high-dimensional data. In particular its ability to capture trends observed in a small training set and to generalize those trends against a broader corpus have made it useful across a large number of applications.

SVM uses a supervised learning approach, which means it learns to classify unseen data based on a set of labeled training data, such as corporate documents. The initial set of training data is typically identified by domain experts and is used to build a model that can be applied to any other data outside the training set. The effort required to construct a high quality training set is quite modest, particularly when compared to the volume of data that may be ultimately classified against it. This means that learning algorithms such as SVM offer an exceptionally cost effective method of text classification for the massive volumes of documents produced by modern organizations. The balance of this paper covers the inner workings of SVM, its application in science and industry, the legal defensibility of the method as well as classification accuracy compared to manual classification.

Overview

SVM is built upon a solid foundation of statistical learning theory. Early classifiers were proposed by Vladimir Vapnik and Alexey Chervonenkis more than 40 years ago. In 1992 Boser, Guyon and Vapnik proposed an improvement that considerably extended the applicability of SVM. From this point on SVM began to establish its reputation as the state-of-the-art method for data categorization. Starting with handwriting recognition tasks SVM showed results that were superior to all other methods of classification. It was quickly shown that SVM was able to beat even Artificial Neural Networks that were considered to be the strongest categorization algorithms at the time. Thousands of researchers applied SVM to a large number of machine learning problems and the results have contributed to the acceptance of this technology as the state-of-the-art for machine classification.

Numerous studies (*Joachmis 1998, Dumais et al. 1998, Drucker et al. 1999*) have shown the superiority of SVM over other machine learning methods for text categorization problems. For example, Joachmis reported 86% accuracy of SVM on classification of the Reuters news dataset, while the next best method, a significantly slower *k*-Nearest-Neighbor algorithm was only able to achieve 82% accuracy.

Today SVM is widely accepted in industry as well as in the academia. For example, Health Discovery Corporation uses SVM in a medical image analysis tool currently licensed to Pfizer, Dow Chemical uses SVM in their research for outlier detection and Reuters uses SVM for text classification.

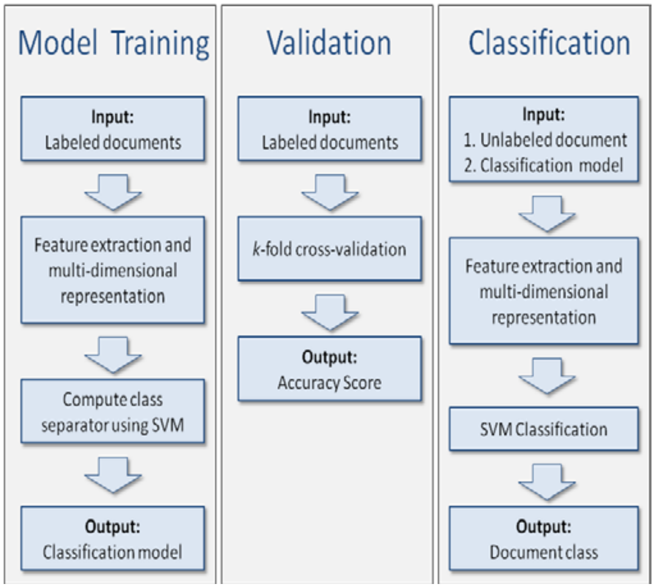

Figure 1: Classification Infrastructure

Under the Hood

Figure 1 shows the typical approach for text classification using SVM. The model is trained using a set of documents labeled by domain experts. The validation procedure computes the expected accuracy of the model on unclassified data. The labeled data itself is used in the accuracy evaluation and therefore the error estimates take into account the specifics of particular data. Once a model is constructed, it can then be used to efficiently and quickly classify new unseen documents in real time.

Model construction

SVM is most commonly used to split a single input set of documents into two distinct subsets. For example, we could classify documents into privileged and non-privileged or record and non-record sets. The SVM algorithm learns to distinguish between the two categories based on a training set of documents that contains labeled examples from both categories. Internally, SVM manipulates documents to represent them as points in a high-dimensional space and then finds a hyper-plane that optimally separates the two categories. In the example on Figure 2, the documents are represented as points in two-dimensional space and the SVM algorithm finds the linear separator (a line) that divides the plot into two parts corresponding to two different classes.

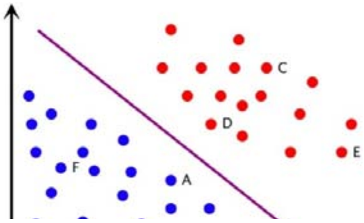

Figure 2: Two dimensional representation of documents in two classes separated by a linear classifier.

The separating line (also called "the model") is recorded and used for classification of new documents. New documents are mapped and classified based on their position with respect to the model. There are many ways to reduce a document to a vector representation that can be used for classification. For example, we could count the number of times particular terms, characters or substrings appeared. We may also consider the lengths of sentences or the amount of white space.

Example: *Say we are trying to classify documents into "work" and "fun" categories. To represent the document as a vector, we could count the number of times the words "meeting" and "play" occurred. In this case the document will be represented as a vector of size two or simply as a point on a two-dimensional plane (like in Figure 2). Similarly, if we count appearance of three or more different words, then the document would be represented in three or more dimensions.*

The dimensions are also called features. There are many kinds of features we can compute on a document including the frequency of appearance of a particular character, the ratio of upper case letters to lower case, the average size of the sentence, etc. Some features are more useful than others, while some are simply noise. Within SVM there exist good algorithms that evaluate how well a particular feature helps in classification. Typically, when documents are prepared for classification the features are extracted, analyzed and the noisiest ones automatically removed.

Once data is preprocessed and a multi-dimensional representation of a document is generated, SVM then finds the optimum hyper-plane to separate the data. As shown in Figure 3 there may be several possible separators, however the SVM algorithm must pick the best one.

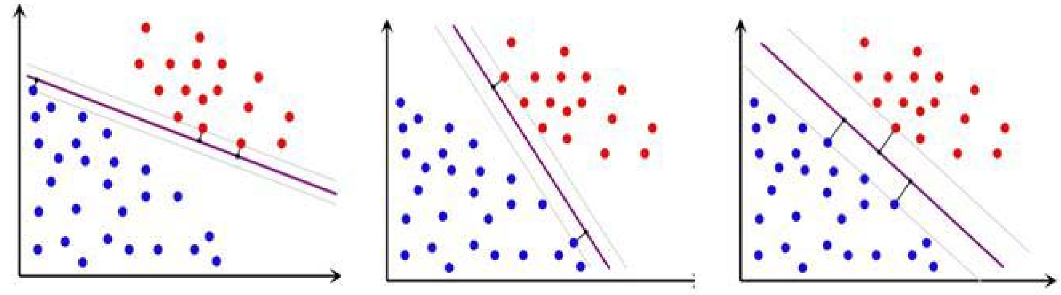

Figure 3 Often there are many possible separators for the data. Support Vector Machines choose the separator with maximum margin as it has best generalization properties.

It is said that one separator is better than another if it generalizes better, i.e. shows better performance on documents outside of the training set. It turns out that the generalization quality of the plane is related to the distance between the plane and the data points that lay on the boundary of the two data classes. These data points are called "support vectors" and the SVM algorithm determines the plane that is as far from all support vectors as possible. In other words SVM finds the separator with a maximum margin and is often called a "maximum margin classifier".

Multi-class classification

The examples above demonstrate classification into two categories; however it is often necessary to group documents into three or more classes. There are established methods of using SVM for multi-class classification. Most commonly an ensemble of binary (two-class) classifiers is used for this problem. In such an ensemble each classifier is trained to recognize one particular category versus all other categories. For each category the classifiers are trained on a labeled set where the documents from the corresponding category make up positive examples and rest of the documents become negative examples. In order to classify new data, each classifier in the ensemble is used to produce a confidence value of a point belonging to the corresponding category and the category with greatest confidence is selected.

Active learning

Much of the effort involved in document classification with SVM comes from the initial review needed for construction of a high quality training set. This effort can be significantly reduced if reviewers' time is used most efficiently - active learning is the best approach to do so. Active learning is a computer assisted method of document review that focuses the reviewers' attention on the areas of the training set that will have the most impact on the training procedure. At first it may be surprising that some documents in the corpus are more important for training then others, but consider the following example. In Figure 2 some documents lay very close to the separator while some are quite far. If we didn't know the labels on the documents far from the separator we could guess the category from their neighborhood. On the other hand, if we didn't know the labels on the documents that are close to the separator and therefore close to documents of the other class, we could not guess the category easily. In fact, the labels on documents which are close to the separator influence the hyper plane much more then labels on the documents that are far from it. Active learning makes use of this observation to make sure that reviewers are always looking at the documents that are close the border line between two classes. By focusing the reviewer's attention to those documents we allow SVM to build the most accurate model while reducing the number of documents needed for training.

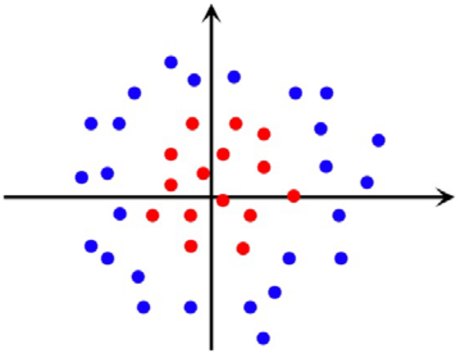

Figure 4: Non-linearly separable data

Non-linear classification

Sometimes the data that we are dealing with is not linearly separable. Figure 4 shows an example of a data set that cannot be separated by any line in two dimensions. SVM has a way to deal with this problem as well. The data may be mathematically projected into higher dimensions where it can be more easily separated. For example, while the data shown in Figure 4 is not linearly separable in two dimensions; it can be separated in three dimensions. To do so, imagine that we loaded this picture into a slide projector and projected it onto a white cone attached to the wall in front of the projector with the tip pointing away from the wall. We center the cone such that its tip is matched with the center of our plot. Then, consider where the points appear on the surface of the cone. All blue points will be projected closer to the tip of the cone and all the red points will appear closer to the other end.

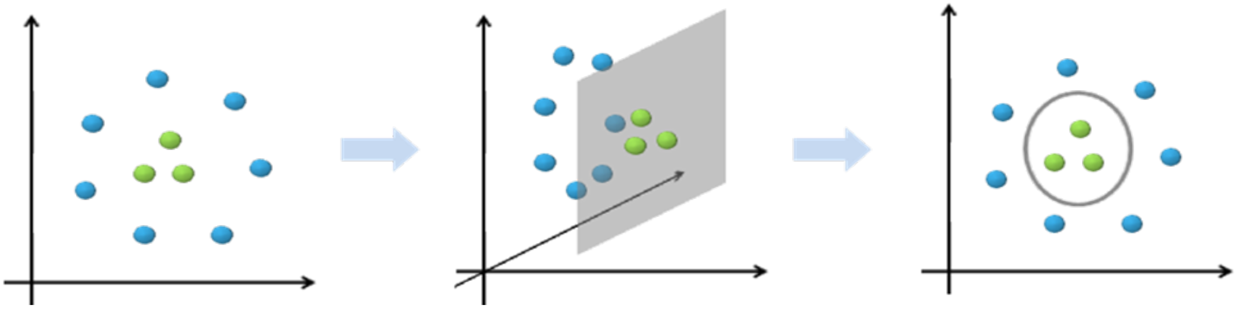

Figure 5: Non-linear classification. First figure shows the dataset that cannon be separated by a linear classifier in two dimensions, but can be projected into three dimensions where it can be separated by the plane.

Now, we can take a plane (a linear separator in three dimensions) and cut the cone, such that all blue points stay on one side of the plane and the green points on the other side. SVM will pick a plane that has good generalization properties to classify unseen data. The cone in this example is called a kernel. In other words, the kernel is the function that describes how to project data into higher dimensions. There are a number of different types of kernels available to fit different data. Good guidelines exist to help in selecting a proper kernel.

Accuracy Evaluation

Understanding the accuracy or the expected rate of success of the algorithm is essential to the successful application in a commercial enterprise. Fortunately solid testing procedures have been developed over the years to evaluate the performance of learning algorithms. *Accuracy* is a measure of how close the results of the automatic classification match the true categories of the documents. Accuracy is estimated by applying the classifier to the *testing dataset* classified by domain experts. The documents for which the classification algorithm and domain experts assigned the same label is said to be classified correctly and the rest of the documents are classified incorrectly. The accuracy is computed as number of correct over number of correct plus number incorrect. In other words, the

$$Accuracy = \frac{Num.Correct}{Num.Correct + Num.Incorrect}$$

accuracy is the percentage of documents that were classified correctly. When evaluating the accuracy it is essential to ensure that documents from the testing set were not used to train the classifier. This ensures that the classifier will have no unfair information about testing documents that will inflate the performance. Typically, the set of all documents labeled by domain experts is split into *training* and *testing sets.* The algorithm is trained using the training set and then applied to the testing set for accuracy evaluation. Since, none of the testing documents appeared in the training set, the performance of the algorithm on the training set will be good estimator of expected accuracy on unseen data.

In order to build up the confidence in the estimated value of accuracy it is beneficial to train the model on multiple training sets and evaluate it against multiple testing sets and then compute the average accuracy. This approach is known as *k*-fold cross-validation and is accomplished by partitioning a single labeled dataset into multiple testing and training sets. As shown in Figure 6, the method prescribes the labeled set to be split into *k* parts, also called folds. Commonly a 10 fold split is considered to be sufficient. For every fold the training set is constructed from *all but one part* of the labeled data. The single remaining part is used as a testing set. This approach results in *k* training/testing sets from the same labeled dataset. The accuracy is evaluated on each of the splits and the average computed.

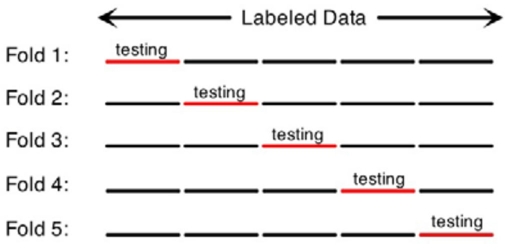

Figure 6: Example of 5-fold cross-validation. Labeled data is split into five parts and for each fold classifier is trained on four parts and validated on one remaining part and the average of fold accuracies is computed.

When making sense of a quality of the classification algorithm it is important to keep in mind that it is typically not possible to reach 100% accuracy, however accuracies of 80%-90% are commonly considered achievable.

To gain the perspective of what 90% accuracy means in the real world it is necessary to compare it to the accuracy of human reviewers. It turns out that on average human reviewers do not perform better than some of the best machine learning algorithms and in many cases humans perform significantly worse. Godbole and Roy, 2008 studied the quality of human classification of natural language texts in the support industry. They found that when different groups of reviewers were asked to review the same set of documents they disagreed on categories for 47% of documents. Furthermore, when the same reviewer was given the same document to review on different occasions their labels only agreed in 64% of cases, this means that the same reviewer did not even agree with themselves 1/3 of the time.

It is now possible to train a machine algorithm that will outperform or work on par with manual classification. Wai Lam et. al., 1999 observed this when comparing the quality of manual and automatic classification of medical literature with respect to text retrieval. Similar observations were reported in the litigation support industry by Anne Kershaw, a founder of nationally recognized litigation management consulting firm. Her group compared the results of automatic and manual privilege coding over population of 48,000 documents and found that automated classification was much more accurate then manual review, minimizing the chance of missing an important privileged document.

Defensibility

To analyze the defensibility of results obtained using SVM classification, consider the related standard for admitting expert scientific testimony in a federal trial. In Daubert vs. Merrell Dow Pharmaceuticals, Mr. Justice Blackman suggested following four factors be considered:

- Whether the theory or technique can be and has been tested
- Whether the theory or technique has been subjected to peer review and publications
- The known or potential rate of error or the existence of standards
- Whether the theory or technique used has been generally accepted

SVM satisfies all four of requirements. The years of research in statistical learning theory as well as thousands of publications that study SVMs completely satisfy the first and second requirements. The extensive testing methodologies available for SVM quantify the expected accuracy of the algorithm and as such completely satisfy the third requirement. The data-centric error rate calculation described in the Accuracy Evaluation section above measures the accuracy of the algorithm as it specifically relates to a particular data. This approach to testing and quality evaluation meets the strictest requirements of modern science. The second and fourth requirements are satisfied by the wide acceptance of SVM as a state-of-the-art method for classification that is broadly utilized in science and industry.

Conclusion

Automatic document classification has become a necessity for any large enterprise. SVM is a mature and well understood technology founded on years of work in statistical learning theory. It is currently the state-of-the-art approach for categorization tasks in a host of settings. It has been studied from a number of perspectives resulting in thousands of publications. It is accurate, scale-able, predictable and auditable; and as such is an ideal match for the burgeoning corporate demand for automatic text and document classification.

Michael McCreary is the Chief Operating Officer of Rational Retention, LLC. Rational Retention is a solution based software company that brings loose enterprise data into compliance with a company's retention policies while providing an immediate ROI by significantly reducing records management and litigation costs. 2 Tower Place, Albany, NY 12208, 518-489-3000.

References

B. Boser, I. Guyon, and V. Vapnik. (1992) A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory.* ACM, Pages: 144 - 152 , 1992

H. Drucker, Donghui Wu, and V. Vapnik. (1999) Support vector machines for spam categorization. *IEEE Transactions on Neural Networks,* IEEE, Pages 1048-1054,

S. Dumais, J. Platt, D. Heckerman and M. Sahami (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of ACM-CIKM98*, Nov. 1998, Pages. 148-155.

T. Joachims (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*. Springer, 1998. Pages 137-142.

S. Godbole and S. Roy, (2008). Text classification, business intelligence, and interactivity: automating C-Sat analysis for services industry. *KDD 2008*: 911-919

W. Lam, M. Ruiz, P. Srinivasan, (1999): Automatic Text categorization and its application to text retrieval. *IEEE Transactions on Knowledge and Data Engineering*

A. Kershaw (2005): Automated Document Review Proves Its Reliability. *Digital Discovery and e-Evidence, Vol. 5, No. 11, November 2005*