Text Categorization and State-of-Art Support Vector Machine

Riddhi Thakkar, Tarjni Vyas
Institute of Technology Nirma University, Ahmedabad, Gujarat, INDIA

Abstract. Task Categorization is task of automatically assign document into their respective class. This paper gives glimpse of text categorization (TC), and explains the principal tasks of a text categorization system. It also list various "Text Categorization" techniques. It mainly focuses on Support Vector Machines (SVMs). Support Vector Machines (SVMs) are the most popular machine learning algorithm used for "Text Categorization", and provides some reasons why SVMs are suitable for this task. This paper also list out extension of SVM, advantages and disadvantages of it. It also list out the applications where SVM is used. It also make comparison between various Text Categorization techniques.

Keywords: Text Categorization, Support Vector Machine (SVM).

1. Introduction

With the fast growth of online information resources, Automatic Text Classification (ATC) has become hot research topic in recent years. Task Categorization is task of automatically assign document into their respective categories. Text Categorization is used to find information on the web, to classify news stories and to filter out spam. Previously Text Categorization is done manually by defining a set of rule that encoding expert knowledge. Now a days Text Categorization is done by using machine learning approach in which the classifier understand rules from examples, then evaluates them on set of test documents.

TC is a supervised learning method. Supervised Learning is process of leaning from pre-given categories and labeled document examples and using that it classify new documents. A supervised learning algorithm is an algorithm which analyzes training data and generates inferred function, which can be used for mapping new example. An optimal scenario which will allow for the algorithm to precisely determine a class labels for unseen instances. This is required by learning algorithm to be generalized from the training data to unseen situations in "reasonable" way.

A linear model is a model which uses the linear composition of feature/values. Positive/negative differentiation is de- pendent on the sign of this linear combination. There are more advanced models neural networks, decision trees etc., however it is very interesting that in case of Text Categorization a linear model is expressive enough to achieve good results. There are lots of linear models: Naive Bayes, support vector machines, K- Nearest Neighbor (KNN), neural networks perceptron. Naive Bayes is very famous among Spam Filters, As it is simple for training and testing. Naive Bayes has optimal training and testing time of O() order and it is proportional to the read through example. Naive Bayes has simplicity to learn from new examples and has ability to modify the existing model. Support Vector Machines have been proven as one of the powerful and popular algorithms for text categorization.

A. Support Vector Machine

SVMs belongs to the family of generalized Linear Classifier and can be interpreted as the extension of Perceptron. They can also be view as a special case of Tikhonov regularization. A special property of Support Vector Machine is that they minimize the actual classification error and maximize the geometric margin at the same time; and so this method is also known as Maximum Margin Classifiers. Now a days Support vector machines are not only better than other machine learning methods, but they are performing at the state-of-the- art level and have ample amount of current theoretical and actual applications. The main Feature of SVM is high accuracy. The implementation and use of Support Vector Machines for classification starts in a similar manner to supervised machine learning problems. A training data set is composed of individual training samples and it is used for producing the classifier. We are given x f, where x is a vector of size. nf demonstrates the quantity of elements. The features are the amount that is used for making a decision. SVM is used to divide space by decision boundary. For any two classes, that are separable by training data sets, there are many possible separators, as shown in figure:

DOI: 10.090592/IJCSC.2016.120

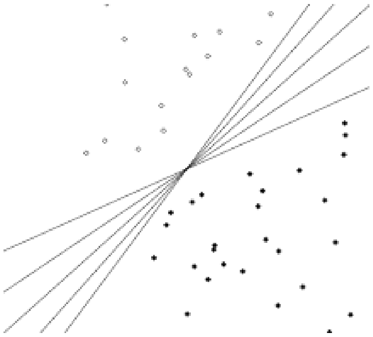


Figure 1 Infinite number of hyper planes that separate two linearly separable classes

Conventionally, a decision boundary drawn in mid of the void between data items of the two classes looks better than. One in which it approaches very close to examples of one or both classes. While some of the supervised learning methods like as a perceptron algorithm that find just any linear separator, some others, such as Naive Bayes, look for the best linear separator with respect to some criterion. The SVM specifically defines the criterion for looking for a decision surface that is maximally far from any data point. This separation from the decision surface to the nearest information point decides the Margin of the classifier. The requirement for construction of this method is that the decision function for SVM IS totally defined by the subset of data which defines the position of separator. And this separator point s are called as Support Vectors. The figure.2 shows the support vector and margin for sample issue.

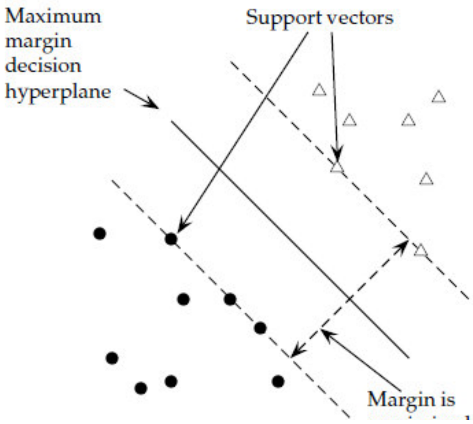


Figure 2 The support vector and margin for sample issue

It is good to maximize the margin because the points near to the decision surface do not represent accurate classification decision, there is almost 50% of chance that it decides in either way. Maximum margin gives safety margin. Safety Margin: A small error in calculation or slight change in document may not cause a misclassification.

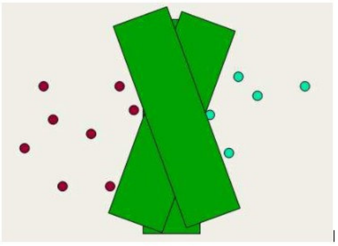


Figure 3 Large Margin Classification

By comparing Decision Hyper plane, we place a large margin between classes, we have less choice for where to put it. And the result is, models memory capacity can be decreased.

B. Formalization of SVM algebra

A decision hyper plane can be defined a decision hyper plane normal vector w that is perpendicular to hyper plane and by intercept term b. w is also refer to as Weight vector in machine learning. b is used to choose all the hyper planes which are perpendicular to normal vector. As the hyper plane is perpendicular to normal vector, all points →−x on the hyper plane satisfy →−w T →−x = −b. For example there is a set of training data points D = (→−w i, →−y i) , where each member is a pair of xi and are labeled by yi class label corresponding to it. In SVM two data classes are labeled as +1 and -1, not by +1 and 0,the intercept term is explicitly represented as b. Now, the Linear Classification is: f(→−x)=sign(→−w T →−x + b) If the classification of a point is far away from the decision boundary, then we are confident about it. Now, for a data set and decision hyper plane, we define the functional margin of the i th example →−x I with respect to a hyper plane ¡→−w , b as a quantity yi = (→−w T →−x i + b) We can always make the functional margin big according to our wish by simply scaling up →−w and b. This recommend that we required to place some constraint on the size of the →−w vector.

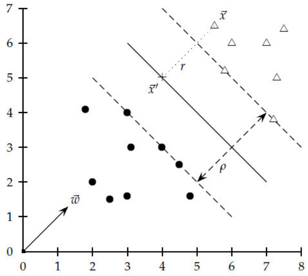


Figure 4 The Geometry margin of Deciding boundary and Point

Now, we estimate Euclidean distance from point →−x to the decision boundary. In figure.4 this distance is denoted by r. The shortest distance between hyper plane and point is perpendicular to the plane, and therefore it is parallel to unit vector. A →−w in this direction is →−w / | →−w | . In the Figure the dotted line is a translation of the vector r →−w / | →−w | . Now Let us label the point as →−x on the hyper plane closest to →−x . Then:

→−x ’ = →−x y r →−w / | →−w |

Here, We multiplying by y just to change the sign for the two cases of →−x as they being on either side of the decision surface. →−x is on the decision boundary and so it satisfies →−w T →−x + b = 0. Therefore,

$$\vec{w}^{\mathrm{T}}\left(\vec{x} - yr\frac{\vec{w}}{|\vec{w}|}\right) + b = 0$$

Solving for r gives,

$$r = y\frac{\vec{w}^{\mathrm{T}}\vec{x} + b}{|\vec{w}|}$$

The geometric margin of the classifier is the maximum width of the band that can be drawn separating the support vectors of the two classes [2]. The geometric margin is clearly independent to variation of parameters: if we change →−w to 6→−w and b to 6b, then the geometric margin is the similar, because it is naturally standardized by the length of →−w . So, it means that we can force any scaling limitation we wish on →−w without influencing the geometric margin. Another choice is that we can use Unit Vector by requiring that | →−w |= 1. This would have the impact of making the geometric margin similar to the functional margin. We can change the functional margin as we wish. So, for all items in the data:

yi(→−w T →−w i + b) ≥ 1

And there exist support vectors for which the inequality is an equality. As each samples distance from hyper plane is ri = yi(→−w T →−w i + b)/ | →−w |, then geometric margin is ρ = 2/ | →−w | . We still want to maximize geometric margin. For, that we want to find →−w and b as such,

DOI: 10.090592/IJCSC.2016.120

ρ = 2/ l →−w l . is maximized,

– For all (→−x i, yi) □ D, yi(→−w T →−w i + b) >=1

Maximizing 2 / l →−w l is similar to minimizing l →−w l /2. From that we get the final standard formula of SVM as a minimization problem:

- 1 / 2 →−w T →−w i
- for all (→−x i, y i), y i (→−w T →−w i + b) ≥ 1

As the constraints and objective function are convex, according to optimization theory, this problem exists the unique uniform minimum solution. We apply Lagrange multiplier to convert it into dual problem:

Max:

$$\sum_{i=1}^{n} y_i a_i = 0, ai \geq 0, i = 1, 2 \cdots n$$

Construction Condition:

Where ai is the Lagrange multiplier for each sample. If is the optimal solution to above equation, then

$$W *= \sum_{i=1}^{n} aiyixi$$

The optimal classification is:

f(x)= sgn{(w□.x) + b□}

= sgn{ $\sum_{i=1}^{n} a *_{*} yi(xi \ \ x)$ + b}

Here, sgn is sign function D.

a. Nonlinear SVM

The method is through a nonlinear mapping to map the sample space to a high-dimensional or even infinite dimensional feature space, so that linear SVM method in the feature space can be applied to solve the nonlinear classification problems in the sample space. The nonlinear mapping from the sample space to the feature space is shown as Fig.5 [3].

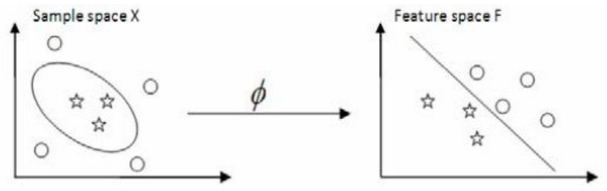


Figure 5 Nonlinear mapping from sample space to feature space

Kernel Function: K (xi, xj) = (xi).(xj) is called kernel function. According to the relevant functional theory [4], if the kernel function satisfies the Mercer condition, it corresponds to the inner product in some transformation space, (x). (xi) = K(x, xi). So, the use of proper kernel function can be an option to non-linear mapping in high dimensional space, to achieve the linear classification. For kernel function K(xi, x), there are three types of SVM.

1. Polynomial kernel: K (x , xj) = (γx.xi + r)d

2. RBF kernel: $K (x , xj) = \exp(-1/2\delta \, \| \, x - xi \;\; 2)$
3. Sigmoid kernel: $K (x , xj) = \tanh[\gamma(x.xi) + r]$

b. The General Process of Text Classification based on SVM

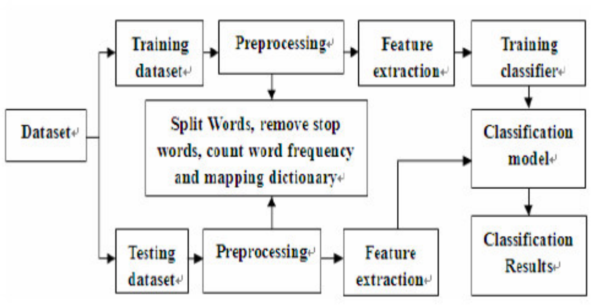


Figure 6: General Text Classification Process

The general process is consisting of mainly four modules: feature extraction, training model, text preprocessing and training classifier.

c. Extensions of SVM

1. Support vector clustering (SVC)
- SVC is a similar method that also builds on kernel functions. But SVC is suitable for data-mining and unsupervised learning.

2. Multiclass SVM
- Aim of Multiclass SVM is to assign the labels to samples by using SVM, where the labels are taken from a finite set of several elements. The principal solution for doing so is to reduce the single multiclass problem into multiple binary classification problems.
- Solution for such reduction include:
- Building binary classifiers
- DAGSVM(Directed acyclic graph SVM)
- Error-correcting output codes

3. Structured SVM
- SVMs have been generalized to structure SVMs. In structured SVMs the label space is structured and it is of possibly infinite size.

d. Joachims Sums up why SVMS are good for Text Categorization
Most text categorization problems are linearly separable.
- High-dimensional input space [6].
- Document vectors are sparse: despite the high dimensionality of the representation, each of the document vectors contain only a few non-zero element [6].
- Few irrelevant features: almost all feature contain considerable information. He conjectures that a good classifier should combine many features and that aggressive feature selection may result in a loss of information.

e. Applications
SVMs can be used to solve various real world problems:
- Hand-written characters can be recognized using SVM.

- SVMs are useful in hypertext and text categorization as their application can notably reduce the requirement for labeled training instances in both the transductive settings and standard inductive.
- Classification of images can also be performed using SVMs. Experimental results show that SVMs achieve significantly higher search accuracy than traditional query refinement schemes after just three to four rounds of relevance feedback.
- The experiments shows that SVM achieves high search ac- curacy than traditional query filtration policy after just some round of relevance feedback.
- SVMs are also helpful in medical science to classify proteins with up to 90% of the composites classified correctly.

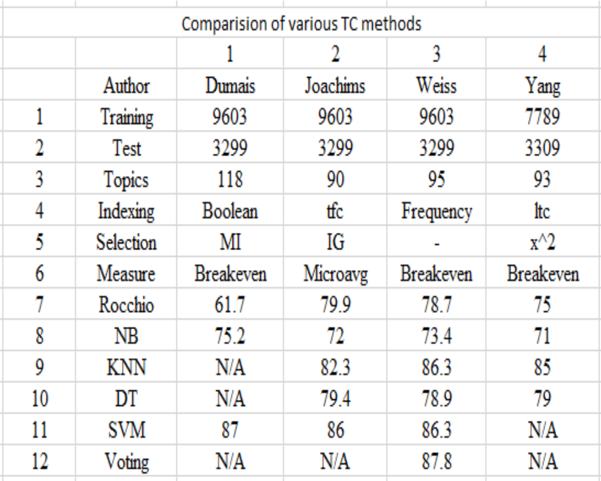
f. Issues of SVM

Potential disadvantages of the SVM are the accompanying three perspectives:

- The main problem of SVM is choice of kernel.
- There is limitation of size and speed in both training and testing.
- Full labeling of input data is required.
- Uncelebrated class enrollment probabilities.
- There is no optimal design for multiclass SVM classifier.
- The SVM is applicable for only two-class tasks. So, algorithms which reduces the multi-class task to several binary problems have to be applied.
- Parameters of solved model are difficult to understand.

g. Comparison of various Text Categorization methods Based on SIX Classifiers

- Classification accuracy: six classifiers (Reuters-21578 collection)
-



Comparision of various TC methods

		1	2	3	4
	Author	Dumais	Joachims	Weiss	Yang
1	Training	9603	9603	9603	7789
2	Test	3299	3299	3299	3309
3	Topics	118	90	95	93
4	Indexing	Boolean	tfc	Frequency	ltc
5	Selection	MI	IG	-	x^2
6	Measure	Breakeven	Microavg	Breakeven	Breakeven
7	Rocchio	61.7	79.9	78.7	75
8	NB	75.2	72	73.4	71
9	KNN	N/A	82.3	86.3	85
10	DT	N/A	79.4	78.9	79
11	SVM	87	86	86.3	N/A
12	Voting	N/A	N/A	87.8	N/A

- From comparison we can say that SVM, Voting and KNN are showing good performance and NB, DT and Rocchio are showing relatively poor performance.

A. Analysis of Result

- Precision is enhanced with an expansion in the quantity of elements until some level. –SVM obtains the best performance.
- None of the algorithm appears to be globally superior over the others; however, SVM is good choices by considering all the factors.

Conclusion

This paper gives some introduction about text categorization, and describes the common tasks of a TC system. Now a days use of Support Vector Machines as a machine learning algorithm is very popular approach. This paper also states some reasons why SVM is very useful for Text Categorization. This paper also include various extension of SVM. It states advantages and Applications of SVM. By Comparing SVM with other TC methods, it conclude that it achieves high performance among others.

Reference

[1] https://en.wikipedia.org/wiki/Supervisedlearning

[2] ”An Introduction to Information Retrieval” , hristopher D. Manning,Prabhakar Raghavan, Hinrich Schtze

[3] ” Study on SVM Compared with the other Text Classification Methods”, Kun Liu,Zhijie Liu,Xueqiang Lv,Shuicai Shi

[4] Research and Application of Support Vector Machine[J](in chinese)., Xiaodan Wang, Jiqin Wang. Research and Application of Support Vec- tor Machine[J](in chinese). Journal of Air Force Engineering Univer- sity(Natural Science Edition), 2004,5(03):49-55.

[5] https://en.wikipedia.org/wiki/Supportvectormachine

[6] Istvn Pilszy, ”Text Categorization and Support Vector Machines,

[7] A. Aizawa ”TAn information-theoretic perspective of tf-idf measures In- formation Processing and Management: an International Journal archive,Vol. 39, Issue 1, 2003, pp. 45-65

[8] Charles Elkan Boosting and Naive Bayesian Learning, Technical Report No. CS97-557,, September 1997, University of California, San Diego

[9] Tom Brey, Larry Lamers, Text categorization with support vector ma- chines:learning with many relevant features,, Proc. of ECML- 98, 10th European Conference on Machine Learning, Springer Verlag, Heidelberg,DE,1998,pp. 137-142

[10] Miao Zhang,De-xian Zhang TTrained SVMs Based Rules Extraction Method for Text Classification

[11] Sundus Hassan,Muhammad Rafi,Muhammad Shahid Shaikh Comparing SVM and NaIve Bayes Classifiers for Text Categorization with Wikitology as knowledge enrichment

[12] Simon Tong,Daphne Koller Support Vector Machine Active Learning with Applications to Text Classification.