IJIRR

International Journal of Information Research and Review
Vol. 03, Issue, 03, pp. 2073-2077 March, 2016

Review Article

INVESTIGATION AND APPLICATION OF IMPROVED TEXT MINING BASED ON SUPPORT VECTOR MACHINE

*Tang Zhi-hang

School of Computer and Communication, Hunan Institute of Engineering Xiangtan 411104, China

ARTICLE INFO

Article History:

Received 19th, December 2015
Received in revised form
05th, January 2016
Accepted 01st, February 2016
Published online 31st, March 2016

Keywords:

Support Vector Machine,
Text mining, Cross-validation.

ABSTRACT

Text mining involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. Text mining can help an organization derive potentially valuable business insights from text-based content. We built a RapidMiner process to examine and learn to classify spam messages. Several thousand messages were analyzed, and a Support Vector Machine learner was able to classify messages with about 91.89% accuracy in a very simple process. We discussed how to examine the frequency of words in documents. The basics of the Support Vector Machine method were explained, as well as cross-validation, and dataset balancing.

INTRODUCTION

Text mining can help an organization derive potentially valuable business insights from text-based content such as word documents, email and postings on social media streams like Face book, Twitter and LinkedIn. Mining unstructured data with natural language processing (NLP), statistical modeling and machine learning techniques can be challenging, however, because natural language text is often inconsistent. It contains ambiguities caused by inconsistent syntax and semantics, including slang, language specific to vertical industries and age groups, double entendres and sarcasm. Text analytics software can help by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques. With an iterative approach, an organization can successfully use text analytics to gain insight into content-specific values such as sentiment, emotion, intensity and relevance. Because text analytics technology is still considered to be an emerging technology, however, results and depth of analysis can vary wildly from vendor to vendor.

Related work

Text Mining (Ananiadou and McNaught, 2006) is the science of leveraging textual data, like whole web pages or narrative fields within a database, for data mining. Text, a type of unstructured data, is challenging due to the richness and complexity of language, but holds enormous potential for forward-thinking firms, due to the sheer volume and depth of available textual data. These opportunities often fall into the categories of Learning from Text or Process Automation. The term text mining describes a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation (Bilisoly, 2008). The latter term is now used more frequently in business settings while "text mining" is used in some of the earliest application areas, dating to the 1980s, notably life-sciences research and government intelligence. The term text analytics also describes that application of text analytics to respond to business problems, whether independently or in conjunction with query and analysis of fielded, numerical data. It is a truism that 80 percent of business-relevant information originates in unstructured form, primarily text. These techniques and processes discover and present knowledge – facts, business rules, and relationships – that is otherwise locked in textual form, impenetrable to automated processing (Sudhahar *et al*., 2015). Traditional learning methods (LIU Wen-kai *et al*., 2008) are mostly based on the theorem of large numbers with the number of samples tends to infinity, but in the actual work, we can't and don't need obtain unlimited number of data. In the application of seismic data for oil and gas projections, the number of samples is often limited, even a small sample of data. Because traditional learning methods are constructed on the

Corresponding author: Tang Zhi-hang,
School of Computer and Communication, Hunan Institute of Engineering Xiangtan 411104, China.

basis of the empirical risk minimization criterion (ERM), therefore, some good learning methods are often less effective in the actual prediction applications with a small sample of data (Maji *et al*., 2008). In the case of a limited sample, although a variety of optimization algorithms are used to minimize the training error, but the learning process will usually result in larger generalization error, so that the model's predictive power to the unknown data is weak, such as the over fitting problem of neural networks had learning problems. To this end, Vapnik proposed criteria for structural risk minimization (SRM) and provided support vector machines (SVM) based on SRM (Bian Zhaoqi *et al*., 2002). Xu and Zhang applied kernel Fisher discriminated to forecast oil and gas reservoirs (Djeffal Abdelhamid *et al*., 2010), which kernel methods are introduced to Fisher discriminated with the emergence of SVM, and achieved good results. Yan, Zhang, *et al*., applied SVM method to sedimentary faces identification (Xu Jianhua *et al*., 2002); the results also demonstrated that in the case of small samples, accuracy of SVM is better than one of neural networks. Support vector machine (SVM) is a kind of generalized linear classifier whose feature is able to simultaneously minimize the empirical classification error and maximize the geometric margin, which are also known as maximum margin classifiers. One special case in classification is the binary classification, and nearly all kinds of the classification algorithm would start their work from binary classification and then to the multiple classification.

Binary classification is frequently performed by using a real-valued function $f: X \subseteq R^n \to R$ in the following way: the input $X = (x_1, x_2, \ldots, x_n)'$ is assigned to the positive class, if $f(x) \geq 0$, and otherwise to the negative class. In the case where $f(x), x \in X$ is a linear function, it can be written as

$$f(x) = < W \cdot X > + b = \sum_{i=1}^{n} w_i x_i + b \tag{1}$$

Where $(W, b) \in R^n \times R$ are the parameters that control the function. The decision rule is given by $\operatorname{sgn}(f(x))$:

$$y = \operatorname{sgn}(f(x)) = \begin{cases} +1 & f(x) \geq 0 \\ -1 & f(x) < 0 \end{cases} \tag{2}$$

The learning process implies that these parameter, W and b, must be learned from the sample data. As shown in Figure 1, it requires not only the optimal classification hyper plane that can separate two categories of sample error-free, but make the distance between the two categories to be maximal. For classification hyper plane to meet this requirement, there is only one. A geometric interpretation of this kind of hypothesis is that the input space X is split into two parts by the hyper plane defined by the equation. A hyper plane is an affine subspace of dimension n-1 to divide the space into two half spaces that correspond to the inputs of the two distinct classes. For example, the hyper plane is the dark line in Figure 1 with the positive region above (y=+1) and the negative region below (y=-1). The vector w that defines a direction perpendicular to the hyper plane is called weight vector, while varying the value of b that moves the hyper plane parallel to itself, is called offset. For specific data, there are numerous in such a hyper plane.

Based on the above discussion, the classifying plane defined by solid thick line is apparently better than the one defined by dashed thick line, because the gap between its two deciding margins is comparatively large. If the gap is too small, any tiny disturbance to the deciding margin would lead to notable influence on the classification. In the other words, the deciding margin which has comparatively large gap has better generalization error than that with small gap. Maximum-margin hyper plane law is about designing the linear classifier for maximum-gap deciding margins to make sure the generalization error reaches its minimal level even under the worst situation. Linear classifier belongs to this kind.

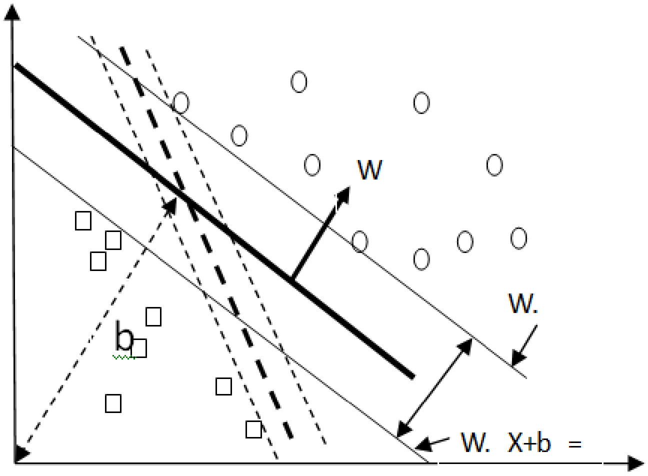


Figure 1. Optimal Classifying Hyper plane

Improved Text mining based on Support Vector Machine

Getting the Data

This paper uses a dataset of 5574 SMS (mobile phone text) messages hosted by the University of California, Irvine, Machine Learning Repository. You can read more about the dataset and download it from here: http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection. It contains 747 messages marked as "spam", and the remainders are non-spam messages marked as "ham". It is a tab-separated text file with one message per line, with UTF-8 encoding, Figure 2 depicts dataset of 5574 SMS and Figure 3 depicts the Meta data view.

Tokenizing the Document

The Tokenize operator takes a document and splits it into a set of words based on the selected mode. In this case, we will split the document on non-letters, meaning that every time the operator encounters a symbol, such as a space or hyphen character, it will split the document into a new token. Thus, "a stitch (in time), saves - nine", would turn into

The tokens {a, stitch, in, time, saves, nine}

1. Select the Process Documents from Data operator, and in the Parameters tab, change vector creation to Term Occurrences. This will let us see the number of times that each word appears in each document.
2. Check keeps text. This will keep the original text in a column in the output table.

3. Using your mouse, drag a line from the Process Documents from Data wor port (for Word List) to a res node on the right of the process.
4. Double-click the Process Documents from Data operator to see its inner process.
5. Find the Tokenize operator in the Operators tab. Double-click the Tokenize operator to add it to the Process Documents from Data inner process. This will automatically connect the correct nodes.
6. Click the blue "up" arrow to exit the inner process.

Examining the Word Vector

A word vector is just a fancy name for a table, where each row is a document (SMS message in this case), and each column is a unique word in the corpus (all of the words in all of your documents). The values inside the table depend on the type of word vector you are creating. In this case we are using Term Occurrences, meaning that a value in a cell represents the number of times that word appeared in that document. You could also use the Binary Term Occurrences, meaning the value in the cell will be zero if the word did not appear in that document, and one if the word appeared one or more times in that document. It is always a good idea to examine your data, in order to "get a feel" for it, and to look for strange anomalies.

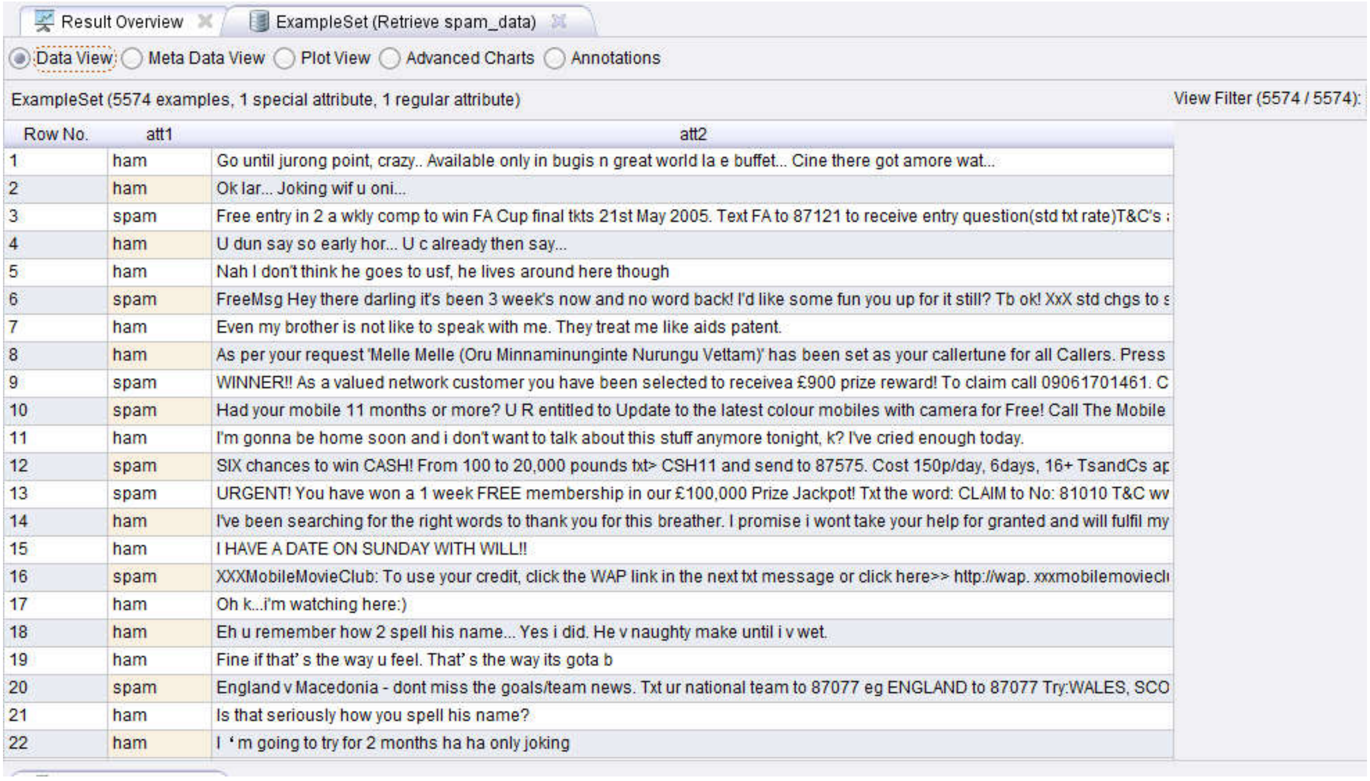
1. Click the Example Set tab to view the word vector. You will see that the word vector has 9755 attributes, meaning that there are 9755 unique words in the corpus. Equivalently, there are 9755 columns in the word vector.
2. Look at the Range column for the "text" role attribute. You will note that:

- "Sorry I'll call later" is the most common message.
- Below that, you can see that there are 4827 "ham" messages, and 747 "spam" messages.

3. Click the Data View button. You will see that document 13 has a "1" under the word "A", meaning that the word "A" appears one time in that document. Actually, "a" and "A" appear in that document, but we are considering the letter case of the words in this process, so they are counted as distinct words, Figure 4 depicts table of wordlist frequencies and Figure 5 depicts table of example set

Validating the Model

To build the Support Vector Machine model, add the Support Vector Machine operator after the Process Documents from Data operator. To find the model's predictive accuracy, we must apply the model to data, and then count how often its

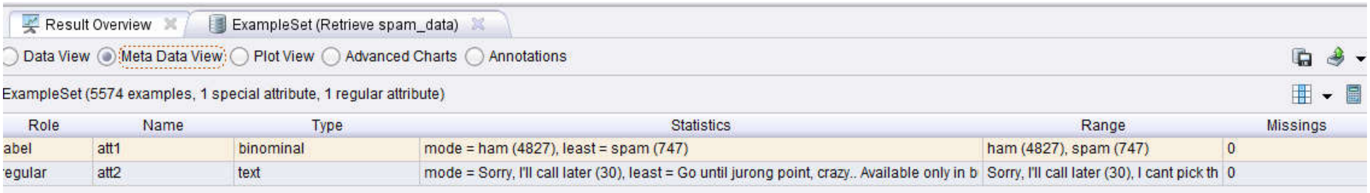

Result Overview
ExampleSet (Retrieve spam_data)
● Data View ○ Meta Data View ○ Plot View ○ Advanced Charts ○ Annotations

ExampleSet (5574 examples, 1 special attribute, 1 regular attribute)

View Filter (5574 / 5574):

Row No.	att1	att2
1	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
2	ham	Ok lar... Joking wif u oni...
3	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's
4	ham	U dun say so early hor... U c already then say...
5	ham	Nah I don't think he goes to usf, he lives around here though
6	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to s
7	ham	Even my brother is not like to speak with me. They treat me like aids patent.
8	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press
9	spam	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. C
10	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile
11	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
12	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs ap
13	spam	URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C ww
14	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my
15	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
16	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemoviecl
17	ham	Oh k...i'm watching here:)
18	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
19	ham	Fine if that' s the way u feel. That' s the way its gota b
20	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCO
21	ham	Is that seriously how you spell his name?
22	ham	I ' m going to try for 2 months ha ha only joking

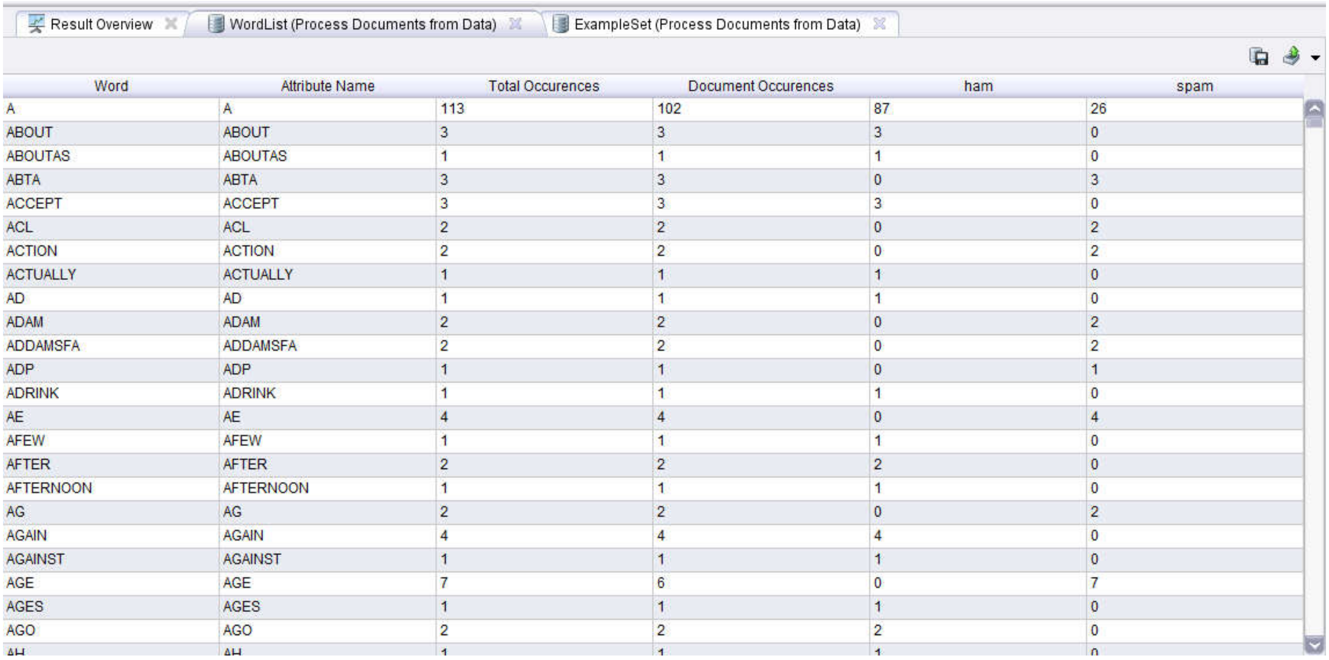
Figure 2. Dataset of 5574 SMS (mobile phone text)


Result Overview
ExampleSet (Retrieve spam_data)
○ Data View ● Meta Data View ○ Plot View ○ Advanced Charts ○ Annotations

ExampleSet (5574 examples, 1 special attribute, 1 regular attribute)

Role	Name	Type	Statistics	Range	Missings
abel	att1	binominal	mode = ham (4827), least = spam (747)	ham (4827), spam (747)	0
egular	att2	text	mode = Sorry, I'll call later (30), least = Go until jurong point, crazy.. Available only in b	Sorry, I'll call later (30), I cant pick th	0

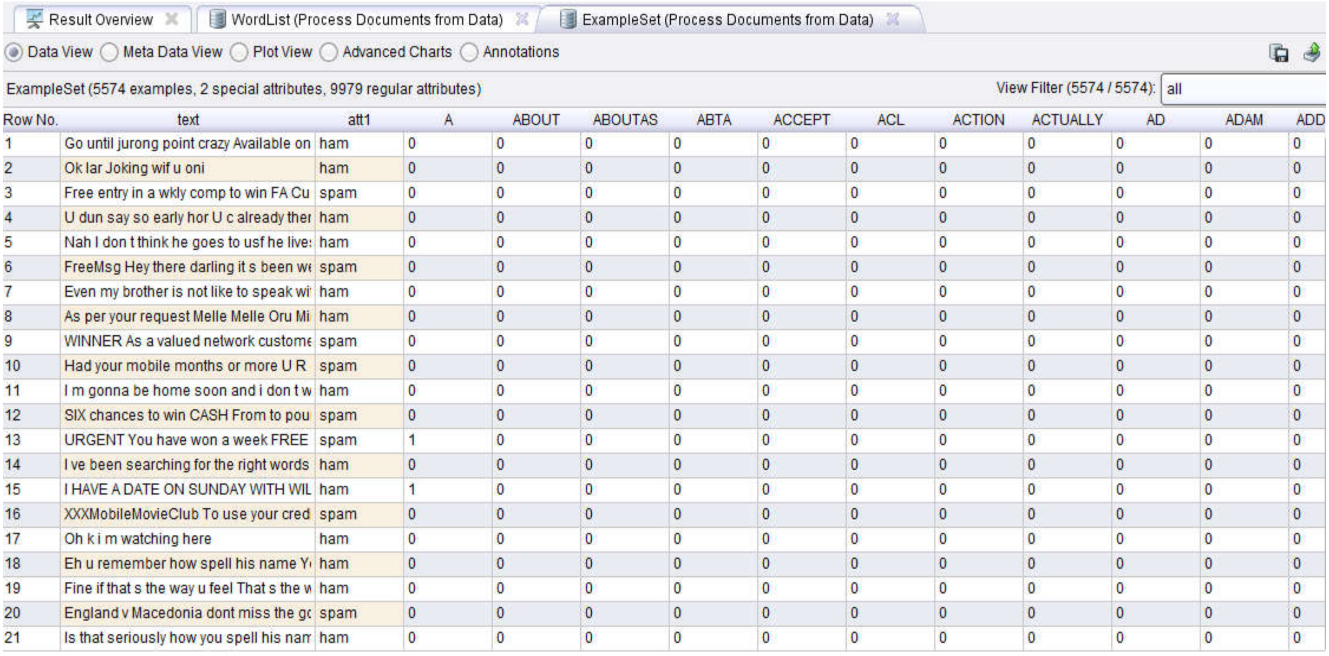
Figure 3. The Meta data view



Result Overview
WordList (Process Documents from Data)
ExampleSet (Process Documents from Data)

Word	Attribute Name	Total Occurences	Document Occurences	ham	spam
A	A	113	102	87	26
ABOUT	ABOUT	3	3	3	0
ABOUTAS	ABOUTAS	1	1	1	0
ABTA	ABTA	3	3	0	3
ACCEPT	ACCEPT	3	3	3	0
ACL	ACL	2	2	0	2
ACTION	ACTION	2	2	0	2
ACTUALLY	ACTUALLY	1	1	1	0
AD	AD	1	1	1	0
ADAM	ADAM	2	2	0	2
ADDAMSFA	ADDAMSFA	2	2	0	2
ADP	ADP	1	1	0	1
ADRINK	ADRINK	1	1	1	0
AE	AE	4	4	0	4
AFEW	AFEW	1	1	1	0
AFTER	AFTER	2	2	2	0
AFTERNOON	AFTERNOON	1	1	1	0
AG	AG	2	2	0	2
AGAIN	AGAIN	4	4	4	0
AGAINST	AGAINST	1	1	1	0
AGE	AGE	7	6	0	7
AGES	AGES	1	1	1	0
AGO	AGO	2	2	2	0
AH	AH	1	1	1	0

Figure 4. A table of wordlist frequencies



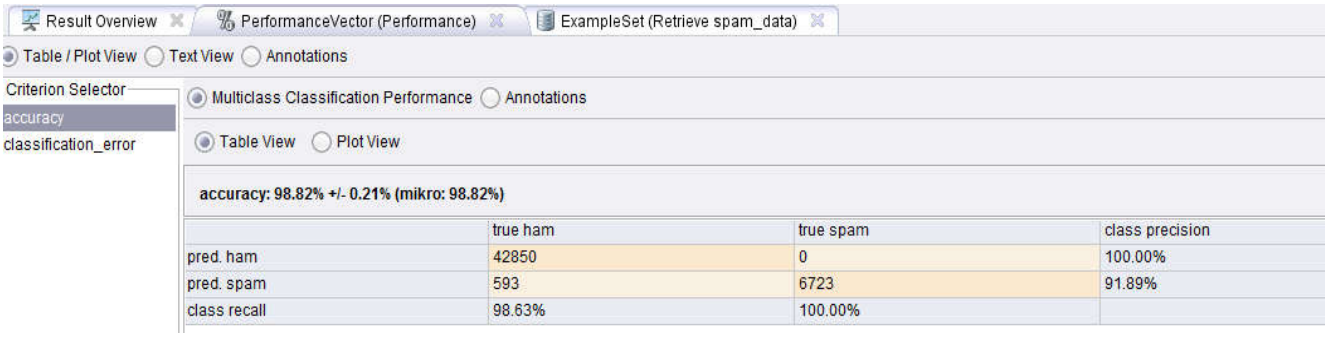
Result Overview
WordList (Process Documents from Data)
ExampleSet (Process Documents from Data)
● Data View ○ Meta Data View ○ Plot View ○ Advanced Charts ○ Annotations

ExampleSet (5574 examples, 2 special attributes, 9979 regular attributes)

View Filter (5574 / 5574): all

Row No.	text	att1	A	ABOUT	ABOUTAS	ABTA	ACCEPT	ACL	ACTION	ACTUALLY	AD	ADAM	ADD
1	Go until jurong point crazy Available on	ham	0	0	0	0	0	0	0	0	0	0	0
2	Ok lar Joking wif u oni	ham	0	0	0	0	0	0	0	0	0	0	0
3	Free entry in a wkly comp to win FA Cu	spam	0	0	0	0	0	0	0	0	0	0	0
4	U dun say so early hor U c already ther	ham	0	0	0	0	0	0	0	0	0	0	0
5	Nah I don t think he goes to usf he live:	ham	0	0	0	0	0	0	0	0	0	0	0
6	FreeMsg Hey there darling it s been we	spam	0	0	0	0	0	0	0	0	0	0	0
7	Even my brother is not like to speak wi	ham	0	0	0	0	0	0	0	0	0	0	0
8	As per your request Melle Melle Oru Mi	ham	0	0	0	0	0	0	0	0	0	0	0
9	WINNER As a valued network custome	spam	0	0	0	0	0	0	0	0	0	0	0
10	Had your mobile months or more U R	spam	0	0	0	0	0	0	0	0	0	0	0
11	I m gonna be home soon and i don t w	ham	0	0	0	0	0	0	0	0	0	0	0
12	SIX chances to win CASH From to pou	spam	0	0	0	0	0	0	0	0	0	0	0
13	URGENT You have won a week FREE	spam	1	0	0	0	0	0	0	0	0	0	0
14	I ve been searching for the right words	ham	0	0	0	0	0	0	0	0	0	0	0
15	I HAVE A DATE ON SUNDAY WITH WIL	ham	1	0	0	0	0	0	0	0	0	0	0
16	XXXMobileMovieClub To use your cred	spam	0	0	0	0	0	0	0	0	0	0	0
17	Oh k i m watching here	ham	0	0	0	0	0	0	0	0	0	0	0
18	Eh u remember how spell his name Y	ham	0	0	0	0	0	0	0	0	0	0	0
19	Fine if that s the way u feel That s the w	ham	0	0	0	0	0	0	0	0	0	0	0
20	England v Macedonia dont miss the gc	spam	0	0	0	0	0	0	0	0	0	0	0
21	Is that seriously how you spell his nar	ham	0	0	0	0	0	0	0	0	0	0	0

Figure 5. A table of example set

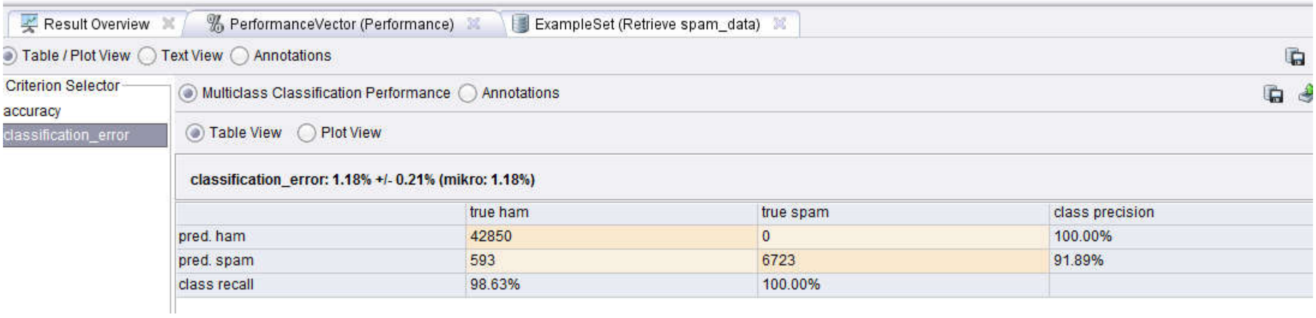


Result Overview
PerformanceVector (Performance)
ExampleSet (Retrieve spam_data)
● Table / Plot View ○ Text View ○ Annotations
Criterion Selector
accuracy
classification_error
● Multiclass Classification Performance ○ Annotations
● Table View ○ Plot View

accuracy: 98.82% +/- 0.21% (mikro: 98.82%)

	true ham	true spam	class precision
pred. ham	42850	0	100.00%
pred. spam	593	6723	91.89%
class recall	98.63%	100.00%	

Figure 6. Accuracy of performance vector


Result Overview
PerformanceVector (Performance)
ExampleSet (Retrieve spam_data)
● Table / Plot View ○ Text View ○ Annotations
Criterion Selector
accuracy
classification_error
● Multiclass Classification Performance ○ Annotations
● Table View ○ Plot View

classification_error: 1.18% +/- 0.21% (mikro: 1.18%)

	true ham	true spam	class precision
pred. ham	42850	0	100.00%
pred. spam	593	6723	91.89%
class recall	98.63%	100.00%	

Figure 7. Classification error of performance vector

predictions are correct. The accuracy of a model is the number of correct predictions out of the total number of predictions. Add the Apply Model operator after the Support Vector Machine operator and connect their two nodes together. Add a Performance operator after the Apply Model operator and connect it to a res node. To be able to predict a model's accuracy on unseen data, we must hide some of the data from the model, and then test the model on that unseen data. One way to do this is to use K-fold Cross-Validation. When using, say, 10-fold Cross-Validation, we would hide 1/10th of the data from the model, build the model on the remaining 9/10ths of the data, and then test the model on the whole dataset, calculating its accuracy. We would do this again, hiding a different 1/10th of the data from the model, and test again. We would do this 10 times in total, and take the average of the accuracies. This provided a better idea of how the model will perform on data that it has not seen before, Figure 6 depicts accuracy of performance vector and Figure 7 depicts classification error of performance vector.

1. Remove the Support Vector Machine, Apply Model, and Performance operators from the Main Process window.
2. Connect an X-Validation operator to the Process Documents from Data operator, and connect its ave (for average performance) node to a res node.
3. Double-click the X-Validation operator. Put a Support Vector Machine operator in the left side of this inner process, and an Apply Model operator and a Performance operator in the right side of the process. Connect all required nodes.

Conclusion

We only looked at one feature of the words in the documents—their frequency. But there are many other features in words that we could also examine. For example, we could examine their length, upper case to lower case ratio, number of symbols, and similar features in the previous and next words. This might lead to a more accurate classifier.

Acknowledgements

A Project Supported by Scientific Research Fund of Hunan Provincial Education Department（15A043）

REFERENCES

Ananiadou, S. and McNaught, J. 2006. Text Mining for Biology and Biomedicine. Artech House Books.

Bilisoly, R. 2008. Practical Text Mining with Perl. New York: John Wiley & Sons.

Sudhahar, S. and Veltri, G.A. 2015. Cristianini. N Automated analysis of the US presidential elections using Big Data and network analysis. Big Data & Society 2 (1), 1-28

LIU Wen-kai, WANG Rui-fang, ZHENG Xiao-juan. 2008. Estimating coal reserves using a support vector machine. *Journal of China University of Mining and Technology,* 18(1), 103-106

Maji, S., Berg, A.C. and Malik, J. 2008. Classification using Intersection Kernel Support Vector Machines is Efficient. *IEEE Computer Vision and Pattern Recogntion*, CVPR 2008, USA, 2008

Bian Zhaoqi, Zhang Xuegong.Pattern Recognition (M). BeiJing: Tsinghua University published,2002

Djeffal Abdelhamid, Babahenini Mohamed Chaouki, Taleb Ahmed Abdelmalik, 2010. An SVM Based System for Automatic Dates Sorting International Review on Computers and Software, 5(4), 423-428

Xu Jianhua, Zhang Xuegong, Li Yanda, 2002. Application of kernel Fisher discriminating technique to prediction of hydrocarbon reservoir (J). OGP, 37(2), 170-174

****