Review for special issue of Computer Speech and Language

Title: Continuous speech recognition using support vector machines Authors: Aravind Ganapathiraju, Jonathan Hamaker and Joseph Picone

Overview

As indicated by the title, this paper is concerned with the application of Support Vector Machines (SVMs) to automatic speech recognition. The first part of the paper is a tutorial review of SVMs, and in particular the Structural Risk Minimisation (SRM) and Expected Risk Minimisation (ERM) training paradigms. The second part of the paper describes the application of SVMs to two problems is speech recognition: first a vowel classification task (the 'Deterding' vowel set), and then a small vocabulary continuous speech recognition task (the OGI alpha-digits). In the second instant, SVMs are used as a secondary classifier using a segmentation of the speech obtained using a conventional HMM-based system. In both experiments the SVM based system achieves superior performance.

Detailed Comments

The paper is not compact and could be shortened substantially without compromising its ability to report the important results.

Comments on SVM 'tutorial' section

The utility of the tutorial section is reduced by a lack of precision and poor notation in the mathematical presentation. For example:

1. (top of page 6): Why is the set of networks from which the optimal network needs to be chosen *finite*?
2. (page 6): What is $\boldsymbol{Z}$? On page 6 you say that it is the *union* of the input vector space and the output vector space. But then (equn (3)) you write $P(z)$, where P is a joint distribution taking values $P(x,y)$, where x and y are elements of the input and output spaces. Surely, therefore, $\boldsymbol{Z}$ is the *cross-product* of the input and output vector spaces.
3. Any function has to have a domain and a range. What is the range of one of the functions g? Similarly, further down the page you talk about the *function* $g(z)$, but $g(z)$ is not a function, g is.
4. The notation in equation (3) is poor. If $\boldsymbol{R}$ is a functional, then it is defined on elements of a function space. You don't say what $g(z)$ is, but I don't think it's a function. I think the left-hand side of (3) should be $R(g)$.
5. This is supposed to be a tutorial, but you don't give any intuitive explanation of what g, L and Q are for, or the properties that they should have. You don't even say what the ranges of any of these functions are (i.e what sets do they take values in?).
6. (top of page 7): What is the range of the sum in (6)?
7. (top of page 7): The interpretation of R_{emp} as the mean error depends on the interpretation of Q, which in turn depends on the interpretation of L. But no intuitive interpretation of these functions is offered.

8. The notation in equations (5) and (7) is inconsistent. In (5) the arguments of Q are pairs of elements of z and parameter sets α. In (7) the arguments of Q are pairs of input and output vectors. Why not use notation such as Q_α to denote a particular instance of Q parameterised by α?
9. The function f has acquired an additional argument in (7) compared with its introduction in section 2.1.
10. (page 7) "Let the minimal actual risk be obtained using the function $Q(z,\alpha_0)$". Surely you mean that the minimal actual risk is obtained with the parameter set α_0.
11. (page 7) In equation (8) why use f again to denote a function which is different from the f in, for example, (7)? Also, please say that h is a scalar value, not just what it is called.
12. (page 8) Equation (10) introduces a new term η, which is not mentioned or described anywhere.
13. (page 9) In bullet point 1 you introduce the notion of 'confidence interval'. This has not been mentioned previously. I assume that the confidence interval is the size of $f(h)$.
14. (page 9) Final paragraph. "SVM learning process optimises for a minimum confidence interval". It is still not clear to me what this means. Does it mean that S the learning process finds a minimum confidence interval? Several of the sentences in this paragraph are ambiguous.
15. (page 10) Line 2. "Figure 3" should be "Figure 4" (same comment later in the paragraph). Also, again you need to be more precise. Earlier in the paragraph you define the margin to be the distance between two hyperplanes. Then you say that the optimal hyperplane is the <u>one</u> that maximises the margin while minimizing the empirical risk This doesn't make sense as you have written it.
16. (page 10) What do you mean by "Let $\boldsymbol{w}$ be a vector that is normal to the decision region". I.e. what does it mean for a vector to be normal to a region?
17. (page 10) I know what an *n-tuple* is, but what is a "tuple". If you mean 2-tuple why not just say '"pair". Also, strictly speaking, $\{x,y\}$ denotes the set whose members are the points x and y. In other words, the order of x and y is not important. If you mean an ordered pair then you should use the notation (x,y).
18. (page 16). You define "margin" twice on this page. First below (11) and then again at the bottom of the page.
19. (page 16) In (12) and (13) why is it obvious that you can write "$\geq+1$" and "≤-1" on the right-hand sides of these equations?
20. (page 11). Therefore, are the support vectors the ones which lie on the separating hyperplanes?
21. (page 11) In equation (15), are the α_is the Lagrange multipliers?
22. (page 11) In equation (18) f is used to denote another, different, function. Why not use a different letter?
23. (page 11) Intuitively, why do the support vectors correspond to non-zero multipliers?

24. (page 13) Second line. “Figure 3” should be “Figure 5”.

Other comments

The terms “neural network” and “multi-layer perceptron” are confused in the paper. In some cases the authors appear to use “neural network” when they mean MLP. Also some of the terminology is rather ‘conversational’. For example, in the second paragraph on page 14, “SVMs….better than other non-linear classifiers like neural networks and mixtures of Gaussians”. There are also several instances of repetition. For example, the assertion that neural networks are not suitable for handling temporal structure is made twice on page 14 – in the first paragraph and in the final paragraph.

Page 15: “…if we assume the effect of $P(A)$ to be insignificant for recognition”. What does this assumption mean?

Page 15: “..between the distance of the test pattern from the margin” Which margin? What does this phrase actually mean?

Page 16: What is the justification for using the sigmoid functions defined in (30)? Also, another use of the letter “f” in equation (31). Also, how good is the Gausian assumption which is made at the bottom of this page?

Page 17: There is probably no need to include equations (32) and (33). However, please explain the sentence after these figures. It appears to confuse the underlying distribution of the data with the distribution which is estimated from the data.

Page 19: In the third paragraph there is a discussion of the roles of the first, second and third states in three state HMM. In my experience (and I believe that of others) it is often the case that such an HMM uses the three states in a parallel, rather than serial, manner. A typical state sequence spends minimal time in two of the three states and most of the HMM occupancy in one state. The state which is occupied for most of the HMM duration varies between speech samples. Thus the behaviour which is described is certainly intuitively plausible but may not always coincide with what actually happens

Comments on experimental section – Deterding vowels

The first experiment which is reported involves classification of log-area parameter frames. More detailed analysis of the results is needed. For example, how many RBF or polynomial kernels were used? Are the RBF width parameters which give the best performance predictable in advance from, for example, the comments in Bishop’s book “Neural networks and pattern recognition”? What types of error are being made, and are they the same as the errors which other types of classifier typically make? Etc.

Comments on experimental section – OGI Alphadigits

The second experiment uses the OGI connected alpha-digit recognition task. The experimental procedure is a two stage process. First the data is segmented using a conventional HMM-based system. Then, for each phone, the three regions corresponding to the initial, second, and final states of a 3 state HMM are processed to yield a 118 dimensional feature vector. Then this feature vector is classified using the SVM classifier. In this way the problem of recognising a time-series is reduced to a sequence of ‘static’ classification tasks.

The horizontal axis of figure 6 should include a scale as well as a label.

It is stated on page 22 that the SVM that for connected word recognition, the SVM model outputs need to be interpreted as probabilities. This is done by observing the SVM output scores for a labelled cross-validation set, and then converting the outputs to probabilities using a similar sigmoid function to that used in section 3.1. This is done "using non-linear optimisation techniques". More precise information would be helpful.

The final recognition result is 11% word error rate, which is the same as that obtained using the HMM system. It is then noted that the patterns of errors for the SVM 'hybrid' and HMM systems are different. Hence an approach which combines both the HMM and SVM methods is proposed. Best performance for this system is 10.6% WER.

This result is certainly of interest. However, I do not believe that the results in their present form are suitable for publication in a journal. The results arise from a sequence of specific decisions: the way in which the speech data is segmented using an HMM, the way in which the SVM is trained, the way in which the SVM outputs are interpreted as probabilities, and the particular use of the segmented data as input to the SVM. There is no analysis of how these components contribute to give the final result. For example, how important is the Structural Risk Minimisation training scheme? How do we know that a similar result would not be achieved using a straightforward Radial Basis Function Network, or some other type of network? How does the HMM segmentation scheme affect the result (it surely cannot be in any sense optimal)? Therefore, although the results are certainly of intriguing, I do not believe that their current presentation is sufficiently deep for a journal article.

Summary

This paper is in two halves. The first is a tutorial on SVM and Structural Risk Minimisation. The second describes a particular way in which SVMs, or an HMM-SVM hybrid system, can be applied to automatic speech recognition.

The first, tutorial , part of the paper is potentially useful, particularly for readers of CSL who are not already familiar with SVMs. However, its usefulness is compromised by poor, and in places inaccurate, mathematical presentation. This part of the paper could also be made substantially shorter, byt removing repetition and more concise presentation.

The results presented in the second part of the paper are undoubtedly of interest. However, in my opinion the presentation is not of sufficient depth, and does not contain sufficient analysis, for journal publication. The paper describes a very specific route to the application of SVMs to automatic speech recognition. Along that route a number of decisions are made which are, arguably, questionable from a speech recognition perspective – for example the use of a conventional HMM-based speech recognition system to segment the OGI data and the conversion of the resulting segments into feature vectors which are suitable for 'static' classification by an SVM. There is no analysis of which parts of the system are responsible for the performance which is reported. For example: Is the structural risk minimisation criterion important, or would the system work just as well with a more conventional criterion? Is the use of an SVM important, or would a more conventional radial basis function network perform just as well? What if the SVM was replaced by an MLP? What would be the effect of replacing the current segmentation procedure with one which was, in some sense, optimal from the perspectives of both the HMM and SVM? For

these reasons I believe that the paper makes a very good conference paper, but is not suitable for publication in a journal such as CSL.

Recommendation

The first part of the paper is not publishable in its present form. The mathematical presentation needs significant improvement and the whole of the section needs shortening, by more concise presentation and the removal of repetition. For journal publication, the second part of the paper also needs major revisions, including the presentation of new experimental results, which explore some of the issues raised in the previous section. In summary, the paper requires major revisions for journal publication.