Multi-class Classification of Cancer Stages from Free-text Histology Reports using Support Vector Machines

Anthony Nguyen, Darren Moore, Iain McCowan and Mary-Jane Courage

***Abstract*— Multi-class machine learning techniques using support vector machines (SVM) are proposed to classify the TNM stage of lung cancer patients from analysis of their free-text histology reports. Stages obtained automatically can be used for retrospective population-level studies of lung cancer outcomes. While the system could in principle be applied to stage different cancer types, the paper focuses on staging lung cancer due to data availability. Experiments have quantified system performance on a corpus of reports from 710 lung cancer patients using four different SVM architectures for multi-class classification. Results show that a system based on standard binary SVM classifiers organised in a hierarchical architecture show the most promise with overall accuracy results of 0.64 and 0.82 across T and N stages, respectively.**

I. Introduction

The TNM cancer staging system is increasingly being recommended as a standard of care by national cancer bodies, e.g. [1]. The preferred mechanism for the TNM staging of lung cancer patients is through a multi-disciplinary team (MDT) conference, which requires input from specialists (e.g. radiology, pathology, etc), and are therefore extremely resource intensive and infeasible in regional areas where local expertise is limited. As a result, formal stage data is not collected for all cancer patients.

Preliminary research towards a decision support system for staging lung cancer patients based on free-text histology reports posed this as a text categorisation problem [2], [3]. A state-of-the-art approach for this is to derive a vector space model document representation and then classify each category using a binary ('one-versus-rest') Support Vector Machine (SVM). An initial system was developed applying binary SVM's to classify each stage category (e.g. T1 versus not-T1) – see [2], [3] for a problem description, literature review and experiments. While promising results were obtained, such a system does not necessarily assign a single stage value within each T and N grouping[1], but rather zero or more. This is a clear limitation as the cancer stage decision is inherently a multi-class problem.

A natural extension of SVM's for multi-class classification is the use of multiple binary classifiers which can be collectively combined in some way to yield a multi-class decision [4], [5], [6], [7]. Some common approaches include comparing each class to all others, and all pairs of classes to each other. Direct methods for training multi-class predictors have also been proposed [6], [8], [9].

This paper addresses the multi-class problem by investigating the performance of four different SVM architectures for multi-class classification with experimental results reported on a corpus of free-text reports from 710 lung cancer patients.

A. Nguyen, D. Moore and I. McCowan are with CSIRO e-Health Research Centre, Brisbane, Australia {Anthony.Nguyen, Darren.Moore, Iain.Mccowan}@csiro.au

M-J. Courage is with the Queensland Cancer Control Analysis Team, Queensland Health, Brisbane, Australia Mary-Jane_Courage@health.qld.gov.au

[1]As metastatic cancer cannot generally be assessed from histological lung examinations, the system does not determine an M stage.

II. Description of System

The proposed system consists of two stages: text pre-processing to standardise report texts, followed by the assignment of the cancer stage via the extraction and classification of features from the report text using SVM-based multi-class classification techniques.

A. Text Pre-processing

The text pre-processing steps implemented in this paper is the same as that described in [3]. For each patient, the input to the system consists of concatenated unstructured text from available histology reports. The steps taken to process the input report text include (1) *normalisation*, which reduces the basic variations between reports by enforcing consistent expression of common terms, (2) *detection of negation phrases*, which detects common medical negation phrases in the text, (3) *conversion to UMLS base forms*, which converts each word in the text document into its Unified Medical Language System (UMLS) SPECIALIST Lexicon [10] base form, (4) *parsing into UMLS term codes*, which converts each base form word into a sequence of more general (i.e. potentially multi-word) UMLS alphanumeric codes, and (5) *negating relevant UMLS terms*, which applies the negation phrases from Step 2 to surrounding UMLS terms that were considered relevant to the TNM lung cancer staging protocol.

B. Assignment of Cancer Stage

The next step in the system is to classify the cancer stage from the pre-processed text using SVM's. A vector space model was used to represent each text document in the data corpus as a vector of term weights $\mathcal{D}_k = (a_{ik})$ (where a_{ik} is the weight of term i in document k). The LTC-weighting scheme as used previously in [3] is again used in this paper.

Four selected SVM-based multi-class classifier architectures (see Fig. 1) will be described in the following section and used in experiments to output a cancer stage decision. The SVM's were implemented using the SVM^{light} toolkit [11], unless otherwise stated.

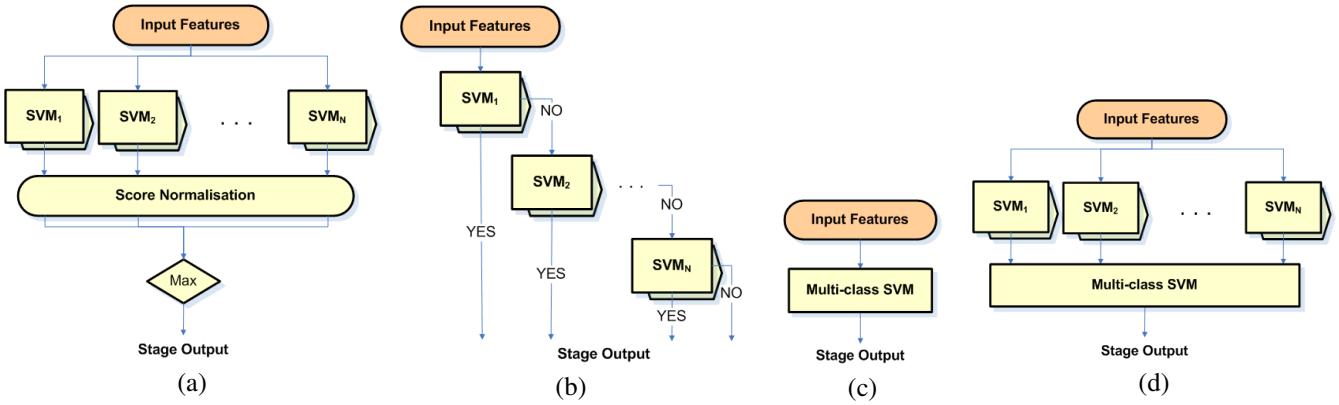

Fig. 1. SVM architectures for multi-class classification, (a) maximum normalised score, (b) hierarchy of binary SVM's, (c) multi-class SVM, and (d) Multi-class SVM on binary SVM outputs.

Maximum normalised score. Binary SVM's output a score indicating the distance of a test example from the decision hyper-plane. These scores are from a classifier-specific distribution and do not directly compare across different SVM's. A sigmoid function fitted to the SVM scores from a development test set [12] was used to normalise the binary SVM scores. The use of the sigmoid function was motivated from empirical observations of real SVM score distributions. By normalising the SVM output scores, a single stage category can be assigned according to the SVM with the highest normalised score.

Hierarchy of binary SVM's. There may be benefit in a hierarchical approach which decomposes the stage classification into a series of sub-decisions (e.g. [7]). With this approach, each classifier performs a pre-defined classification subtask. For example, to classify a T stage (T[1-4]), three binary classifiers need to be trained, say T1 vs. T2/T3/T4, T2 vs. T3/T4, and T3 vs. T4. Permutations of the individual classifiers can be trained to determine the hierarchical structure which best discriminates between the classes.

Multi-class SVM. Multi-class SVM algorithms (e.g. [9]) allow the training of an SVM that does multi-class classification directly. All classes are considered at the same time, and the separating hyper-plane conditions are integrated into a single optimisation problem. The complexity of this approach has variables proportional to the number of categories, and in general, it is computationally more expensive than multiple binary classifiers that use the same amount of data [6]. The SVM^{multiclass} toolkit [13] was used to train multi-class SVM's.

Multi-class SVM on binary SVM outputs. This method uses the binary SVM output scores as inputs to a subsequent Multi-class SVM. The use of both un-normalised and normalised binary SVM scores were considered. Input vectors consisting of a concatenation of both types of scores were found experimentally to yield the best performance.

TABLE I
KEY STATISTICS FOR THE EXPERIMENTAL DATA SET.

Stage	Data	Cases	Reports
T1, T2, T3, T4	Histology + TNM	710	817
N0, N1, N2	Histology + TNM	651	756

III. EXPERIMENTAL EVALUATION

A. Data Corpus

To train and validate the system, a corpus of de-identified medical reports with corresponding staging data was obtained for 718 lung cancer patients following research ethics approval. The corpus was compiled from two separate data sources: a database of pathologic staging decisions (TNM) for lung cancer patients (Queensland Integrated Lung Cancer Outcomes Project data [14]) for use as the gold standard for the classifier training and testing, and a set of histology reports for lung cancer patients extracted from the Queensland Health Pathology Information System (AUSLAB). Only a single histology report was retrieved for most patients.

Only a few T0 (representing patients who did not actually have any tumour) and no N3 (surgical resection is rarely conducted for such cases) cases were present in the source data and so these categories were removed from the experiments. The system also does not classify the TX and NX stage categories. In practice, relevance classifiers can be used to first classify these stages where insufficient information is available for T and N staging. Table I summarises the key statistics for this experimental data set.

B. SVM Training and Testing

The SVM classifiers in the following experiments use a linear kernel with parameters estimated from a training corpus of text reports supplemented with stage (ground truth) data. In experiments not reported here, it was found that different SVM kernel functions (e.g. linear, polynomial, RBF) had negligible effect on the task at hand.

In order to maximise the amount of SVM training data while still reporting significant results on this limited sized

data set, cross-validation using 100 folds, similar to that used in [3], was applied. Within the training subsets, further cross-validation using 10 folds was used to optimise SVM training hyper-parameters. These optimal parameters were used to train a final classifier on the training data subsets and then results were generated on the test subset. As such, over the 100 folds, unbiased results were able to be reported on the full patient list.

C. Performance Measures

The measure of system performance for multi-class classifiers is commonly given by *accuracy*, which is the proportion of correct classifications across all categories and patients. System performance at a per-class level, however, is reported as *recall*, which is the proportion of categories that were correctly assigned by the classifier (true positives / (true positives + false negatives)). In general, the overall accuracy results tend to be dominated by classifier performance on the most common categories.

A confusion matrix can also be used to highlight commonly occurring class confusions. It is a two-dimensional table of frequency counts according to classified (test) class labels and actual (gold standard) class labels. For a perfect classifier, all off-diagonal entries should be zero. This is a useful tool for analysing multi-class classification systems.

Mutual information (MI) can also be derived from a normalised confusion matrix (where all entries sum to 1) to measure the reduction in uncertainty about the ground truth due to the knowledge of the classifier output [15], [16], as given by $MI(y;t) = \sum_y \sum_t P(y,t)\log_2(P(y,t)/(P(y)P(t)))$, where $P(y,t)$ is the joint probability distribution function of the classifier output, y, and the ground truth, t, and $P(y)$ and $P(t)$ are the marginal probability distribution functions of the classifier output and ground truth, respectively. Note that $P(y,t)\log_2(\cdot)$ is defined as 0 if $P(y,t) = 0$. If there is no dependence between the classifier output and ground truth, then the two are statistically independent and by definition their MI is zero. If, on the other hand, the classifier output and ground truth were strongly related, then the value of MI would be relatively high. The magnitude of MI can thus be used to compare different classification systems that use the same data. An evaluation of commonly used metrics in [16] cautioned the use of accuracy, precision, recall and F-score when comparing classifiers as they can be misleading and inconsistent with our intuition about the characteristics of a good classifier. On the contrary, mutual information was shown to better correlate with intuition and able to rank classifiers according to how informative their output was.

IV. Results and Discussion

The T and N staging recall and overall accuracy results for each multi-class method are shown in Table II. In terms of overall system performance, the accuracy results were as high as 0.65 and 0.82 across the T and N categories respectively, which is promising for these initial multi-class experiments. The lower T result, reflect the higher degree of subjectivity and thus difficulty with T staging, as revealed in a concurrent trial involving two clinical experts. It should be noted that although the accuracy results for all multi-class systems are comparable, the result can be misleading due to the dominance of the most common categories.

TABLE III
Confusion matrices for the Hierarchical SVM classifier.

		System Output				
		T1	T2	T3	T4	Σ
Ground Truth	T1	**107**	94	1	2	204
	T2	93	**284**	9	19	405
	T3	1	10	**38**	3	52
	T4	5	13	5	**26**	49
	Σ	206	401	53	50	710

		System Output			
		N0	N1	N2	Σ
Ground Truth	N0	**394**	32	4	430
	N1	35	**91**	23	149
	N2	4	20	**48**	72
	Σ	433	143	75	651

Mutual information, on the other hand, shows that for both the T and N categories, the Hierarchical SVM method is the best classifier in terms of having the most mutual information between the classifier output and ground truth. That is the method has the least uncertainty about the ground truth given knowledge about the classifier output. A closer inspection of the results at a per-class level reveals that this is indeed the case:

1) Maximum normalised score and both Multi-class SVM methods tend to favour one classifier output over others.
2) Multi-class SVM methods are not significantly better than simpler methods that use binary SVM's. While Multi-class SVM on binary SVM's performs best (or close to best) in terms of accuracy, the method is not significantly better than the Hierarchical SVM method when compared at a per-class level.
3) Hierarchical SVM's show promising results both at an overall system level and even more so at a per-class level. It is superior to every other method on all per-class categories except for the most common category.

Given that the Hierarchical SVM method uses binary SVM's, which results in lower computational complexity than methods using Multi-class SVM's, and that they have the potential flexibility of optimising their decision point at each level of the hierarchy (via the use of thresholds on SVM scores), the method shows the most promise in terms of performance, flexibility and simplicity.

The corresponding confusion matrices for the Hierarchical SVM classifiers showing the break-down of cases by stage are shown in Table III. The confusion matrices show that the most common confusions are between T1 and T2 (187 cases), and T2 and T4 (32 cases). As for the N stage confusion matrix, most of the N stage errors are either false positive (52 cases) or false negative (58 cases) findings of N1.

TABLE II

MULTI-CLASS CLASSIFIER RECALL, ACCURACY AND MUTUAL INFORMATION RESULTS.

	Stage	Positive Cases	Max Norm Score	Hierarchical[a]	Multi-class SVM	Multi-class SVM on Binary SVMs[b]
T	T1	204	0.397	**0.525**	0.485	0.338
	T2	405	0.852	0.701	0.765	**0.857**
	T3	52	0.288	**0.731**	0.538	0.635
	T4	49	0.224	**0.531**	0.163	0.265
	Overall Accuracy	710	0.637	0.641	0.627	**0.651**
	Mutual Information		0.123	**0.240**	0.153	0.183
N	N0	430	0.958	0.916	**0.963**	0.960
	N1	149	0.517	**0.611**	0.544	0.503
	N2	72	0.486	**0.667**	0.458	0.569
	Overall Accuracy	651	0.806	**0.819**	0.811	0.813
	Mutual Information		0.295	**0.344**	0.297	0.325

[a]T stage configuration: T3 vs T1/T2/T4, T1 vs T2/T4 and T2 vs T4; N stage configuration: N2 vs N0/N1 and N0 vs N1.
[b]Both normalised and un-normalised SVM scores were concatenated to form the SVM^{multiclass} input.

V. FUTURE WORK

The work discussed in this paper is based on classifying cancer stages from a vector space model representation of patient text reports. The results show that there are obvious class confusions, but it is difficult to determine the reasons for the confusions, thus limiting the utility of the cancer staging tool. For this reason, there may be benefit in indirectly classifying staging factors first at a sentence-level, and then merging the staging factor results using the staging guidelines. That is, individual observations within reports that are relevant to staging are detected, e.g. "primary tumour greater than 3cm in extent". It is envisaged that not only improved performances may result, but also the system will be of greater utility as a tool for retrospective collection of population stage data since it provides the utility for the analysis, extraction and linking of key information from medical reports.

VI. CONCLUSION

Progress towards a system to assist in the collection of staging data for lung cancer patients has been presented. Four multi-class SVM-based classifiers were evaluated. Results show that a system based on standard binary SVM classifiers organised in a hierarchical architecture shows the most promise in our context. The method achieves overall accuracy results of 0.64 and 0.82 across T and N stages respectively for pathologic staging based on histology report text. While this is an interesting initial result, there is much scope to improve the system to incorporate specific knowledge of the staging protocol and practices.

ACKNOWLEDGEMENTS

This research was done in partnership with the Queensland Cancer Control and Analysis Team (QCCAT). In particular, the authors wish to acknowledge: Hazel Harden, Shoni Colquist and Steven Armstrong from QCCAT for their help in system concept definition, and for providing access to data and clinical experts; Jaccalyne Brady and Donna Fry from QILCOP for explaining cancer stage decision making processes; Dr Rayleen Bowman and Dr Belinda Clarke for their expert advice; and Wayne Watson from the AUSLAB Support Group for extracting histology reports to form the research data corpus.

REFERENCES

[1] *Clinical practice guidelines for the prevention, diagnosis and management of lung cancer*. The Cancer Council Australia, 2004.

[2] I. McCowan, D. Moore, and M-J. Fry. Automated cancer stage classification from free-text histology reports. In *Proceedings of the Health Informatics Conference*, Sydney, 2006.

[3] I. McCowan, D. Moore, and M-J. Fry. Classification of cancer stage from free-text histology reports. In *Proceedings of the IEEE Engineering in Medicine and Biology Conference*, 2006.

[4] E.L. Allwein, R.E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

[5] K. Duan and S. Keerthi. Which is the best multiclass svm method? an empirical study. In *Multiple Classifier Systems*, pages 278–285, 2005.

[6] C. Hsu and C. Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, March 2002.

[7] Friedhelm Schwenker. Hierarchical support vector machines for multi-class pattern recognition. In *Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, pages 561–565, Brighton, UK, 30 Aug - 1 Sept 2000.

[8] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[9] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector learning for interdependent and structured output spaces. In *Proceedings of the International Conference on Machine Learning*, 2004.

[10] NIH. Unified Medical Language System (UMLS), 2006. http://www.nlm.nih.gov/research/umls/.

[11] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

[12] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*. MIT Press, 1999.

[13] T Joachims. http://svmlight.joachims.org/svm_multiclass.html.

[14] K.M. Fong, R. Bowman, D. Fielding, R. Abraham, M. Windsor M, and G. Pratt. Queensland integrated lung cancer outcomes project (QILCOP): Initial accrual and preliminary data from the first 30 months. Abstract of Presentation at The Thoracic Society of Australia and New Zealand Annual Scientific Meeting, April 2003.

[15] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[16] H Wallach. Evaluation metrics for hard classifiers. Unpublished note (http://www.inference.phy.cam.ac.uk/hmw26/papers/evaluation.ps), 2004.