Dublin Institute of Technology
ARROW@DIT

Conference papers | School of Computing

2012

An Investigation into Feature Selection for Oncological Survival Prediction

Dmitry Strunkin
Kazan State University, dstrunkin@yandex.ru

Brian Mac Namee
Dublin Institute of Technology

John Kelleher
Dublin Institute of Technology, john.d.kelleher@dit.ie

Follow this and additional works at: http://arrow.dit.ie/scschcomcon

 Part of the Computer Sciences Commons

Recommended Citation

Strunkin, D., Mac Namee, B. & Kelleher, J.D. (2012) An Investigation into Feature Selection of Oncological Survival Prediction. *Ninth International Conference Information Technology: New Generations (ITNG)*, Las Vegas, Nevada 16-18 April. doi:10.1109/ITNG.2012.148

This Conference Paper is brought to you for free and open access by the School of Computing at ARROW@DIT. It has been accepted for inclusion in Conference papers by an authorized administrator of ARROW@DIT. For more information, please contact yvonne.desmond@dit.ie, arrow.admin@dit.ie, brian.widdis@dit.ie.

An Investigation into Feature Selection for Oncological Survival Prediction

Dmitry Strunkin
Applied Mathematics and Informatics Department
Kazan State Technical University
Kazan, Russia
Email: dstrunkin@yandex.ru

Brian Mac Namee
Applied Intelligence Research Centre
Dublin Institute of Technology
Dublin, Ireland
Email: brian.macnamee@dit.ie

John D. Kelleher
Applied Intelligence Research Centre
Dublin Institute of Technology
Dublin, Ireland
Email: john.d.kelleher@dit.ie

***Abstract*—In machine learning based clinical decision support (CDS) systems the features used to train prediction models are of paramount importance. Strong features will lead to accurate models, whereas as weak features will have the opposite effect. Feature sets can either be designed by domain experts, or automatically extracted for unstructured data that happens to be available from some process other than a CDS system. This paper compares the usefulness of structured expert-designed features to features extracted from unstructured data sources in an oncological survival prediction application scenario.**

I. Introduction

Clinical decision support (CDS) systems [1] are a particularly promising real-life application of machine learning techniques. Good CDS systems need to strike a delicate balance between giving medical practitioners confidence in system reliability and taking advantage of the kind of insight that the automated analysis of large multivariate data sources can allow - insight that is not available to typical medical practitioners.

One of the ways in which medical practitioners can be given more confidence in CDS systems is to involve them closely in their design. For predictive machine learning systems the design of input feature sets is one of the key influences that practitioners can exact on system design. Often, however, in certain application domains there is an opportunity to access further secondary unstructured data with which to augment these hand-crafted feature sets with the expectation that this secondary data may enhance the predictive power of resulting models. This, unfortunately, is done at the risk of actually damaging the predictive power of the resulting models [2] and so we must ask whether unstructured secondary data is worth including in predictive models or whether it is better to focus on well understood features designed by experts?

This paper examines this question. Using an oncological survival prediction problem we compare the performance of survival prediction models built using a set of features designed by expert oncologists with the performance of prediction models built using these designed features augmented with features automatically extracted from secondary unstructured data sources relating to the same patients. The data in question is in the form of discharge reports collected over a ten year period in a real oncology clinic.

The paper continues with a review of medical CDS systems before discussing in detail the experiments performed. After presenting the results of these experiments we conclude with a discussion of the implications of these results.

II. Background

The potential of computer science to play a useful role in clinical decision support (CDS) has been recognised since the early 1960s [3]. Many of the early CDS systems were built using handcrafted rules elicited from medical practitioners; the parade example of this type of work being the MYCIN system for diagnosis of blood infections [4]. Although this hand-crafted approach has had success these systems suffered from the fact that they required large time inputs from medical experts. However, at the same time as these systems were being developed work also began on machine learning based approaches to CDS system development.

Machine learning systems are developed by automatically deriving predictive classification models from large data sets. As such they are well suited to the development of medical predictive classification models as medical datasets recording patient symptoms and correct diagnoses are a byproduct of standard hospital recording procedures. One of the earliest systems to use an automatically derived model was the Leeds Abdominal Pain System [5]. This systems used Bayesian probability theory and a large patient dataset to derive a model that could calculate the probability of seven possible causes of acute abdominal pain. Building on this early success a strong tradition of probability based CDS systems has developed, for examples see: [6], [7], [8].

With the maturation of other machine learning techniques, such as decision tree induction [9] and back propogation neural net training [10], the range of machine learning approaches applied to the development of CDS systems broadened - for example neural net based approaches (e.g., [11], [12], [13]) and decision tree based approaches (e.g., [14], [15]).

However, no matter which machine learning algorithm is applied a crucial issue is the design of the feature set used for training the model. Including all possible information can result in the algorithm overfitting the data with a concomitant reduction in the generalizability of the resulting model [2].

Conversely, excluding features from the data set may deprive the algorithm of information that it could use to increase the predictive power of models that are built. Furthermore, the question of whether to use only structured data specifically designed for training a CDS system, or to use all available data (for example unstructured free-text fields) is important. The focus of this paper is an investigation of the use of such secondary unstructured data in a survival prediction problem for oncology patients.

III. Experiments

To investigate the usefulness of different kinds of feature sets in building survival prediction models we performed a series of experiments on an oncological problem. This section will describe, in detail, the nature of these experiments beginning with a description of the dataset used and how this data was manipulated; then covering the experimental methodology used before presenting the results of the experiments.

A. Datasets

All of the experiments described in this paper use a dataset from the Urology Department of the Republican Clinical Oncologic Dispensary in Kazan in the Russian Federation. This dataset contains discharge reports for patients diagnosed with kidney cancer that were treated in the clinic between 2000 and 2010. The overall dataset contains 843 fully anonymised records.

Each discharge report contains 25 structured data points describing the state of a patient's health; the results of medical tests and other analyses; the history of a patient's contacts with the clinic; whether or not the patient survived their cancer; and information on the treatment of the patient. Information is presented in the form of numeric values (e.g. patient age), categorical values (e.g. operation type) and three free-text fields that describe (1) a patient's medical history, (2) a description of any surgical treatments carried out and (3) a description of any diagnostics carried out. These free-text fields were present on the form used by oncologists to make reports, and they were free to make whatever entries they chose. Many oncologists choose not to use these fields, while others use them frequently. Similarly, the length of the texts entered by oncologists vary from just a few words, to in-depth descriptions made up of hundreds of words. Regardless of their length the texts, which are in the Russian language, typically contain various abbreviations, medical terminology and differing typography. Some examples of free-text entries made by oncologists are shown in Figure 1.

На РКТ доброкачественная опухоль?

УЗИ+РКТ - киста.

На УЗИ не выявили, только на КТ: 3 опухолевых очага в почке и подозрение на метастазы в лимфоузлы.

Н/полюс 7х8 см. Подозрение на тромб.

Ныряющий тромб

РКТ подозрение на опухоль

5,5 см

Fig. 1. Examples of text from the free-text fields in the oncology dataset

B. Data Preparation

Using the raw dataset three derived target variable were generated indicating whether or not a patient had survived 1, 3 and 5 years after treatment respectively. So that prediction modules could be developed for each of these targets the full dataset was divided into three subsets in which the total time that a patient had been under observation by the clinic exceeded the survival time being predicted. The sizes of these datasets and the distribution of target variables within them are shown in Table I.

In order to fully explore the potential contribution that the free-text fields could make to prediction model building, two approaches were taken to extracting meaningful data from them: (1) automated extraction and (2) manual extraction. In the automated extraction approach a *bag-of-words* feature extraction (in which a document is represented as a vector of the words it contains from a possible vocabulary) was used. Simple whitespace tokenisation was used and feature reduction was achieved by removing terms that appeared in only 3 or fewer documents. Stop word removal, using a Russian stop word list, was also used. This approach resulted in different numbers of features for each of the three datasets, as shown in Table I.

The second feature extraction approach used was to manually extract medically interesting features from the free-text fields. The features extracted and their levels were:

- disease relapse - boolean
- presence of metastasis[1] - boolean
- presence of metastasis to an adrenal gland - boolean
- presence of metastasis to bones - boolean
- the size of a tumor - numeric
- presence of thrombi[2] - boolean

Using a set of instructions given by an oncologist, student volunteers were used to extract these features if present within any of the free-text fields in the dataset. In cases where the text fields were empty, or where no reference was made to the particular feature a default of false for boolean values and 0 for numeric values was used.

C. Experiment Methodology

Th purpose of the experiments undertaken was to compare the performance of classifiers built using different combina-

[1]metastasis refers to the spread of cancer cells from the initial or primary site of the disease to another part of the body

[2]thrombi are clots in blood vessels

TABLE I
THE DETAILS OF THE THREE SMALLER DATASETS GENERATED FROM THE OVERALL ONCOLOGY SET

Name	Size	Survived Class Size	Died Class Size	Bag of Words Features
Patients under observation for $>$ 1 year	841	741	100	103
Patients under observation for $>$ 3 year	564	443	121	65
Patients under observation for $>$ 5 year	342	254	88	23

tions of features. The combinations considered were:

1) The designed feature set only
2) The features automatically extracted from the text fields only
3) The features manually extracted from the text fields only
4) The designed feature set plus the features automatically extracted from the text fields
5) The designed feature set plus the features manually extracted from the text fields

After some initial experimentation with different classification algorithms three approaches were considered for the presentation of results: (1) the naive Bayes classifier [16], the Voting Feature Intervals (VFI) [17] classifier and the J48 Decision Tree classifier [18]. These classification approaches were chosen because the naive Bayes and decision tree classifers are widely used in CDS systems and the VFI clasisifer achieved particularly good preformenace on the tasks being investigated. Implementations of these classification approaches from the Weka platform [19] were used in all experiments. For each of the three classification problems outlined in Section III-B a 10-fold cross validation experiment was performed for each of the 5 feature combinations described above.

D. Results

The results from the different 10-fold cross validation experiments are shown in Tables II - IV. In these tables we present overall classification accuracy, the accuracies achieved on each of the two classes (*survived* and *died*) and the average class accuracy - a more informative measure than overall classification accuracy for datasets in which the class distribution is imbalanced. For each dataset and classifier type the best average class accuracy achieved is highlighted in bold.

The first point to note about these results is that in general the prediction accuracies of the system are not especially high. This is not unexpected as this sort of classification problem is particularly difficult and mid-range accuracies are common - this is one of the reasons why these sort of systems are only used as decision support tools and final decisions are left to a human expert.

There is also a considerable amount of variation in the performance of the different classification algorithms. It was not possible to build a decision tree classifier using the three datasets that performed well. As is particularly evident from the Died Class Accuracy scores in Tables II and IV the decision tree models built simply could not distinguish between the two classes. The VFI classifier, on the other hand was the best predictive model we could create and outperformed the other two classification approaches in almost all cases.

It is worth noting that the models built using only the features extracted from the free-text fields (both manually and automatically extracted) do not perform well. This suggests that the although there is enough information in these feature sets to inform a prediction, the predictive power of these sets alone is not especially high. Similarly, as the time horizon for which survival prediction is being performed is increased, the accuracy of predictions decreases. Again this is not unexpected as the underlying classification problem becomes more difficult as the time horizon is stretched further into the future.

Finally, in all but three cases (the prediction for survival after 1 year using a naive Bayes classifier and the predictions for survival after 1 and 5 years using a decision tree classifier), the addition of features from the free-text fields has a negative impact on classification accuracy. The times wher this is not the case are when the underlying models themselves are performing poorly. This result addresses the question that is at the core of this paper. In this application scenario there appears to be no benefit in adding extra unstructured information to the structured feature set designed by expert oncologists.

IV. CONCLUSIONS AND FUTURE WORK

The purpose of this paper is to examine the usefulness of secondary free-text data as well as structured, expert-designed features in building models for an oncological survival prediction problem. Typically, survival prediction models use hand-designed feature sets that are the result of input from domain experts, in this case oncologists. While it intuitively makes sense that hand-designed feature sets would lead to good prediction models, and the involvement of domain experts in the design and development of decision support systems is useful in instilling confidence in domain experts in the potential of CDS systems, there is often an opportunity to add extra predictive power to systems by adding model features that may not be obvious to domain experts. These features can often be mined from secondary sources not directly designed for the task.

In the scenario considered in this paper oncologists had the opportunity to add free-text descriptions to the structured data collected in patient discharge reports designed to be used for survival prediction in an oncology clinic. The question asked in the experimental section of this paper is whether or not features extracted from these free-text fields could be useful in developing more predictive models than those built simply using the feature set hand-designed by the collaborating oncologists. Interestingly, in this case, it appears that this extra information is not useful for building prediction models.

TABLE II
SURVIVAL PREDICTION 10-FOLD CROSS VALIDATION ACCURACIES FOR PATIENTS UNDER OBSERVATION FOR >1 YEAR

		Manually extracted features only	Automatically extracted features only	Designed features only	Designed and manually extracted features	Designed and automatically extracted features
Naive Bayes	Overall Accuracy	0.869	0.816	0.872	0.872	0.850
	Died Class Accuracy	0.12	0.337	0.53	0.56	0.64
	Survived Class Accuracy	0.97	0.881	0.918	0.914	0.879
	Average Class Accuracy	0.545	0.609	0.720	0.737	**0.7595**
Voting Feature Intervals	Overall Accuracy	0.830	0.876	0.838	0.836	0.835
	Died Class Accuracy	0.36	0.04	0.62	0.59	0.57
	Survived Class Accuracy	0.893	0.991	0.868	0.869	0.87
	Average Class Accuracy	0.6265	0.5155	**0.764**	0.7295	0.72
Decision Tree	Overall Accuracy	0.881	0.875	0.869	0.862	0.879
	Died Class Accuracy	0	0.03	0.08	0.09	0.13
	Survived Class Accuracy	1	0.991	0.976	0.966	0.98
	Average Class Accuracy	0.5	0.5105	0.528	0.528	**0.555**

Performance better than that achieved using the designed set of features could not be achieved with either extra features manually extracted from the free-text fields in the discharge reports, or from sets of features automatically extracted from the free-text fields. While this is only a small initial experiment on a single application scenario it is indicative of the power of including domain experts *in-the-loop* when developing decision support systems, and reinforces the fact that the design of feature sets is a particularly useful phase in which to solicit expert involvement.

To continue this work in the future we intend to perform larger studies using more datasets with similar characteristics to the data used in the experiments described in this paper. We also intend to investigate the use of interactive data understanding tools, such as those described in [20], to aid domain experts in the design of useful features.

REFERENCES

[1] B. ES, "Clinical decision support systems: State of the art." gency for Healthcare Research and Quality, Tech. Rep. AHRQ Publication No. 09-0069-EF, 2009.

[2] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[3] R. Ledley and L. Lusted, "Reasoning foundations of medical diagnosis," *Science*, vol. 130, pp. 9–21, 1959.

[4] E. Shortliffe, *Computer Based Medical Consultations: MYCIN*. Elsevier/North-Holland, 1976.

[5] F. D. Dombal, D. Leaper, J. Staniland, A. McCann, and J. Horrocks, "Computer-aided diagnosis of acute abdominal pain," *British Medical Journal*, vol. 2, pp. 9–13, 1972.

[6] E. R. I. Kononenko, I. Bratko, "Esperiments in automatic learning of medical diagnostic rules," in *International School for the Synthesis of Expert's Knowledge Workshop*, Bled, Slovenia, August 1984.

[7] I. Kononenko, "Inductive and bayesian learning in medical diagnosis," *Applied Artificial Intelligence*, vol. 7, pp. 317–337, 1993.

[8] G. Lindgaard, P. Egan, C. Jones, C. Pyper, M. Frize, R. Walker, C. Boutilier, B. Hui, S. Narasimhan, J. Folkens, and B. Winogron, "Intelligent decision support in medicine: back to bayes?" *Journal of Universal Computer Science*, vol. 14, no. 16, pp. 2720–2736, aug 2008.

[9] J. Quinlan, *Expert Systems in the Microelectronic Age*. Edinburgh University Press, 1979, ch. Discovering rules from large collections of examples.

[10] R. W. D.E. Rumelhart, G.E. Hinton, *Parallell Distributed Processing*. Cambridge: MIT Press, 1986, vol. 1, ch. Learning internal representations by error propogation.

[11] F. J.J. and D. K.J., "Artificial neural networks for decision support in clinical medicine," *Annals of Medicine*, vol. 27, no. 5, pp. 509–517, 1995.

[12] D. West and V. West, "Model selection for a medical diagnostic decision support system: a breast cancer detection case," *Artificial Intelligence in Medicine*, vol. 20, no. 3, pp. 183 – 204, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365700000634

[13] P. J. Lisboa and A. F. Taktak, "The use of artificial neural networks in decision support in cancer: A systematic review," *Neural Networks*, vol. 19, no. 4, pp. 408 – 415, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608005002844

[14] B. Cestnik, I. Kononenko, and I. Bratko, "Assistant 86: A knowledge elicitation tool for sophisticated users," in *Proceedings of EWSL-87*, Bled, Yugoslavia, 1987.

[15] T. J., B. M., S. L., M. J., and R.-P. A., "Stratification of the severity of critically ill patients with classification trees," *BMC Medical Research Methodology*, p. 83, 2009.

[16] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, 1995, pp. 338–345.

[17] G. Demiröz and H. Güvenir, "Classification by voting feature intervals," in *Proceedings of 9th European Conference on Machine Learning*. Prague, Czech Republic: Springer-Verlag, LNAI 1224, 1997, pp. 85–92.

[18] J. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[20] B. Mac Namee and S. J. Delany, "Cbtv: Visualising case bases for similarity measure design and selection," in *Proceedings of the International Conference on Case-based Reasoning (ICCBR) 2010*, 2010.

TABLE III
SURVIVAL PREDICTION 10-FOLD CROSS VALIDATION ACCURACIES FOR PATIENTS UNDER OBSERVATION FOR >3 YEARS

		Manually extracted features only	Automatically extracted features only	Designed features only	Designed and manually extracted features	Designed and automatically extracted features
Naive Bayes	Overall Accuracy	0.789	0.745	0.828	0.817	0.793
	Died Class Accuracy	0.157	0.248	0.512	0.521	0.5
	Survived Class Accuracy	0.962	0.88	0.914	0.898	0.872
	Average Class Accuracy	0.5595	0.564	**0.713**	0.7095	0.686
Voting Feature Intervals	Overall Accuracy	0.766	0.789	0.819	0.814	0.810
	Died Class Accuracy	0.248	0.041	0.587	0.595	0.567
	Survived Class Accuracy	0.907	0.993	0.883	0.874	0.876
	Average Class Accuracy	0.5775	0.517	**0.735**	0.7345	0.7215
Decision Tree	Overall Accuracy	0.796	0.782	0.817	0.810	0.817
	Died Class Accuracy	0.116	0.066	0.331	0.298	0.308
	Survived Class Accuracy	0.982	0.977	0.95	0.95	0.955
	Average Class Accuracy	0.549	0.5215	**0.6405**	0.624	0.6315

TABLE IV
SURVIVAL PREDICTION 10-FOLD CROSS VALIDATION ACCURACIES FOR PATIENTS UNDER OBSERVATION FOR >5 YEARS

		Manually extracted features only	Automatically extracted features only	Initial designed features only	Initial designed and manually extracted features	Initial designed and automatically extracted features
Naive Bayes	Overall Accuracy	0.725	0.737	0.789	0.772	0.763
	Died Class Accuracy	0.125	0.148	0.443	0.409	0.42
	Survived Class Accuracy	0.933	0.941	0.909	0.898	0.882
	Average Class Accuracy	0.529	0.5445	**0.676**	0.6535	0.655
Voting Feature Intervals	Overall Accuracy	0.740	0.754	0.769	0.754	0.763
	Died Class Accuracy	0.182	0.08	0.545	0.511	0.523
	Survived Class Accuracy	0.933	0.988	0.846	0.839	0.846
	Average Class Accuracy	0.5575	0.534	**0.6955**	0.675	0.6845
Decision Tree	Overall Accuracy	0.734	0.743	0.725	0.734	0.725
	Died Class Accuracy	0.091	0	0.034	0.091	0.023
	Survived Class Accuracy	0.957	1	0.965	0.957	0.969
	Average Class Accuracy	**0.524**	0.5	0.4995	**0.524**	0.496