Exploring Associations between Changes in Ambient Temperature and Stroke Occurrence: Comparative Analysis using Global and Personalised Modelling Methods

Wen Liang, Yingjie Hu, Nikola Kasabov, and Valery Feigin

Knowledge Engineering and Discovery Research Institute,
Auckland University of Technology, New Zealand
{linda.liang,raphael.hu,nikola.kasabov,valery.feigin}@aut.ac.nz

Abstract. Stroke is a major cause of disability and mortality in most economically developed countries that increasing global importance. Up till now, there is uncertainty regarding the effect of weather conditions on stoke occurrence. This paper is offering a comparative study of exploring associations between changes in ambient temperature and stroke occurrence using global and personalised modelling methods. Our study has explored weather conditions have significant impact on stroke occurrence. In addition, our experimental results show that the personalised modelling approach outperforms the global modelling approach.

Key words: weather; stoke occurrence; personalised modelling; global modelling; FaLK-SVM.

1 Introduction

Stroke is known as an acute cerebrovascular disease (CVD), it can cause neurological damages or even death (particularly in the elderly) by the reason of the blood supply suddenly disrupted or stopped to part of the brain. It is becoming a major public health concern and challenge in many countries.

Recently, there is increasing evidence linking weather conditions and stroke occurrence. However, thus far, there only few studies on exploring the effect of weather on stroke occurrence, which remains a matter of uncertainty and controversy. From early evidence, environmental triggers of different stroke subtypes are dependent to age, gender and climate characteristics. However, these data are selection bias (e.g. unclear CT/MRI verification of different stroke subtypes), or no reliable data exists in various population groups (e.g. by age, gender, and region).

To date, various technologies have been adopted to study complex stroke data, normally divided into two categories including statistical methods and machine learning methods (e.g. conventional statistical methods are more widely applied, in particular). However, in many cases, the conventional statistical methods have limitations in efficiency and improving the prediction accuracy compared to machine learning methods. Khnsla et al. (2010) presented an integrated machine learning approach that significantly outperformed the Cox proportional hazards model (one of the most popular used statistical methods in medical research) on the Cardiovascular Health Study (CHS) dataset for stroke risk prediction.

Personalised modelling is an emerging machine learning approach, where a model is created for every single new input vector of the problem space based on its nearest neighbours using transductive reasoning approach [8]. The basic philosophy behind this approach when applied to medicine is that every person is different from others, thus he/she needs and deserves a personalised model and treatment that best predicts possible outcomes for this person. Such characteristic makes personalised modelling an appropriate method for solving complex modelling problems.

This paper therefore presents a comparative analysis using global and personalised modelling methods to explore associations between changes in ambient temperature and stroke occurrence. This knowledge will contribute to the understanding of environmental triggers of stroke. In turn, this will help identify other new areas of research, such as physiological studies on weather-stroke associations or clinical trials, to test preventive strategies to reduce the hazardous effects of harmful weather conditions.

The remainder of this paper is organized as follows. Section 2 briefly reviews global and personalised modelling methods. Section 3 describes a recently developed personalised modelling method, Fast Local Kernel Support Vector Machines (FaLK-SVM), that outperforms the global and traditional personalised modelling methods. Section 4 provides the experimental results of the comparative study. Finally, section 5 gives the conclusion and future direction.

2 Background and Related Work

2.1 Global Modelling

A global model is created from the entire data set for the whole problem space based on the inductive inference method. It focuses on the whole problem space rather than individual vectors. This model is usually difficult to be adapted on new incoming input vectors.

Support vector machine (SVM) is one of popular global modelling algorithm, which has been widely used to deal with regression and classification problems. It is a supervised learning algorithm based on small-sample Statistical Learning Theory, which was originally proposed by Vapnik (1998) and his co-workers. It is a powerful tool for separating a set of binary labeled data in a feature space by an optimal hyperplane. The two major types of SVM used far and wide, are linear SVM (Vapnik & Lerner, 1963) and non-linear SVM (Aizerman & Braverman, 1964).

2.2 Personalised Modelling

In contrast to the global modelling, a personalised model is created for every single new input vector of the problem space based on its nearest neighbours using the transductive reasoning approach (Kasabov, 2007). It is more concerned with solving an individual given problem rather than solving a general problem.

Personalised modelling has been successfully applied to deal with a variety of modelling problems. For instance, in personalised healthcare, the knowledge discovered by this approach has clearly shown the contribution to the prediction, diagnosis and therapy for individual patients' diseases (Iakovidis, 2007; Baek et al., n.d.). Furthermore, given the current advances in networking technologies, personalised mobile service offers a more efficient service, which in turn also benefits business (Lankhorst, Kranenburg, Salden, & Peddemors, 2002).

Nowadays, personalised medicine is becoming a leading trend in medicine, health care and life science. As presented by Lesko (2007) who is from the U.S. Food and Drug Administration, "Personalized medicine can be viewed ... as a comprehensive, prospective approach to preventing, diagnosing, treating, and monitoring disease in ways that achieve optimal individual health-care decisions." Ginsburg and McCarthy (2001) presents the objective of personalised medicine to determine a patient's disease at the molecular level, so the right therapies are able to be applied on the right people at the right time. Multiple examples have significantly proved that the traditional form of medicine is declining in favour of more accurate marker-assisted diagnosis and treatment.

K-nearest neighbour (KNN) is the simplest personalised modelling algorithm, was originally proposed by Fix and Hodges in 1951. It is a supervised learning algorithm that has been successfully used for classifying sets of samples based on nearest training samples in a multi-dimensional feature space by using some suitable distance metric such as Euclidean distance or Manhattan distance.

Weighted K-Nearest Neighbour (WKNN) is a successful extension of KNN as developed by Dudani in 1976. It has been widely used to evaluate the output of a model focusing on solely an individual point of a problem space using information related to this point (Vapnik, 1998).

In the WKNN algorithm, each single vector requires a local model that is able to best fit each new input vector rather than a global model, thus each new input vector can be matched to an individual model without taking any specific information about existing vectors into account.

In contrast to KNN, the output of a new input vector is calculated not only dependent upon its k-nearest neighbour vectors, but also upon the distance between the existing vectors and the new input vector which is represented as a weight vector w , this being the basic idea behind the WKNN algorithm.

3 Fast Local Kernel Support Vector Machines - A Recently Developed Personalised Algorithm

Fast Local Kernel Support Vector Machines (FaLK-SVM) is a fast and scalable local SVM algorithm (Segata & Blanzieri, 2010). It trains a set of local SVMs on redundant neighbourhoods in the training set and use for prediction a model which is the nearest to each testing point.

FaLKM-lib implements the Cover Tree data-structure (Beygelzimer, Kakade & Langford, 2006) for fast retrieval of neighbourhoods in the feature space, and integrates the LibSVM for SVM training and prediction (Chang & Lin, 2001).

For clarity, the pseudo code of FalKM-lib algorithm is briefly summarised in Algorithm 1.

Algorithm 1 FaLK-SVM algorithm

```
**Training Stage:**
$model() \leftarrow null$ // the set of candidate personalised models
$model_{point}() \leftarrow null$ // the set of pointers to the models
$c \leftarrow 0$ // the counter for the centres of the models
$idx()$ $\{1, ..., tN\}$ // the indices for selecting centres
initialise $idx$ randomly
**for** $i = 1 \rightarrow tN$ **do**
    $idx_a \leftarrow idx(i)$ // get the $i^{th}$ index
    **if** $model_{point}() =$ null **then**
        $localPoints() \leftarrow$ get ordered kNN of $D_{trn}(i)$ // retrieve its kNN
        $model(c) \leftarrow$ SVMtrain on $localPoints()$ // train a local SVM
        $model_{point}(idx_a) \leftarrow model(c)$ // assign the centre to the trained model
        **for** $j = 1 \rightarrow k'$ **do**
            $idx_b \leftarrow$ get index of $localPoints(j)$
            **if** $model_{point}(idx_b) =$ null **then**
                $model_{point}(idx_b) \leftarrow model(c)$ // to the $c^{th}$ model
            **end if**
        **end for**
        $c \leftarrow c + 1$
    **end if**
**end for**
**Output** $models, model_{point}$

**Testing Stage:**
Set $x_p \leftarrow$ the nearest training point of $x_q$ from $D_{trn}$
Set $idx_p \leftarrow$ index of $x_p$ // retrieve the index of $x_p$
**Output:** $y_q = f(model_{point}(idx_p), x_q)$ // use the corresponding model for predict
    the label of the testing point

***where:***
    $D_{trn}$ is the training set, $tN$ is the training size, $k$ is the neighbourhood size, $k$' is
    the assignment neighbourhood size, $x_p \in D_{trn}$, $x_q$ is a testing data point, $y_q$ is the
    result of $x_q$, $f$ is a prediction function.
```

4 Experiment

4.1 Dataset and Pre-processing

The Weather and Stroke Occurrence dataset is the first time used as a case study to explore the significant associations between changes in ambient temperature and stroke occurrence. This international collaborative study is carried out under the auspices of six population regions, for people over the age of 15: Auckland (NZ), Perth and Melbourne (Australia), Oxfordshire (UK), Dijon (France), and Norrbotten and Vasterbotten counties (Northern Sweden).

The dataset consists of 11,453 samples (all with first-ever occurrence of stroke) and 9 features (4 patient clinical features - categorical data & 5 weather features - continuous data). Individual patient data includes information such as age, gender, history of hypertension and smoking status. Weather data included information such as temperature, humidity, wind speed, windchill and atmospheric pressure. All these weather parameters are measured for the date of stroke occurrence.

We applied *case crossover* design for the data pre-processing, because there is no "non-stroke" patients in the original dataset. We use the day of stroke occurrence as the "stroke" group and 30 days before stroke occurrence for the same participant as the "normal/control" group, assuming that weather parameters 30 days before the index stroke had no influence on the stroke occurrence 30 days later. This approach is known as *case-crossover* design. Mukamel and his colleagues [12] adopted case-crossover design for comparing the measures of weather and ambient air pollution on the day of presentation and control days for each patient.

In this work, we count down 30 days from the date of stroke occurrence. For instance, in order to do the analysis for the patient who firstly trigged stroke on *1 Jan, 1981*, we used the weather measurement data collected 30 days before, i.e. *1 Dec 1980*. If the patient does not have the weather measurement data on 1 Dec 1980, we used the next following stroke occurrence (e.g. 1 Dec 1982 or 1 Dec 1983) measurement data. This approach is based on the assumption that the weather does not have big changes for the same month but in different year.

4.2 Experiment Setup

In this study, the associations between changes in ambient temperature and stroke occurrence were explored only for the Auckland region. This region originally contains 2850 samples, we randomly selected 500 samples to be explored as a preliminary study. So the data set used for the experiments contains 1000 samples (500 "normal/control" patients (class1) and 500 "stroke" patients (class 2)).

We set up the experiments in two ways: (1) using all 9 features to perform a comparative analysis of the global and personalised modelling approaches; (2)using 6 features (age and 5 weather parameters) due to they are all continuous data. So we normalised the data using the linear normalisation method.

Cross-validation is defined as an optimal method for splitting/sampling data. In this study, K-fold (3-folds) cross-validation was used to evaluate the performance of global and personalised modelling approaches. The entire data set is divided into 3 equal-sized subsets. Each time, an individual subset serves as the testing data for testing the model, while the remaining 2 subsets serve as training data. The process of cross-validation is repeated by 3 times with each of subset being estimated exactly once as the testing data. Once all samples have been estimated, the overall accuracy is calculated as the average accuracy across all 3 times experiments.

4.3 Experiment Result and Discussion

A comparison of classification performance using 9 features and 6 features from SVM, KNN, WKNN and FaLK-SVM are summarized in Table 1 & 2, respectively.

As shown in Table 1, the best accuracy is manifested by the FaLK-SVM personalised model, trained on all 9 features. Its classification accuracy is 51.69% (46.81% for class 1 - Normal, and 56.75% for class 2 - Stroke). FaLK-SVM provides better results compared with the traditional SVM, KNN and WKNN methods. However, the accuracy is very low, which is not good enough to be used for exploring the association between weather conditions and stoke occurrence. The reason of cause producing unpromising accuracy might be the weather variables are continuous data, whereas the personal variables are categorical data. It is difficult to study these two types of data at the same time by using different models. Therefore, in the second experiment, we decide to use 6 features (all continuous data) to see whether have any improvement (see results in Table 2).

Table 2 demonstrates the best classification accuracy is achieved by the FaLK-SVM personalised model. Its accuracy is 70.27% (69.90% for class 1 - Normal, and 70.64% for class 2 - Stroke). This compares with the 64.60% accuracy of the SVM model, 65.70% of the KNN model and 65.80% of the WKNN model. It can be obviously seen that the overall classification accuracy, also the class 1 and class 2 accuracy are significantly improved as compared using 9 features. The reason is the data is normalised using the linear normalisation method before the experiment. As a general conclusion, we could say that weather conditions have significant impact on stroke occurrence. From the experiment, we also find out that all the models use same amount of features but provide different accuracy. So the k-nearest neighbour could be an important factor to improve the accuracy, which worth to be further studied in the future.

5 Conclusion and Future Direction

In this study, we present a comparative study of exploring associations between changes in ambient temperature and stroke occurrence using global and personalised modelling methods. Our study has explored weather conditions have significant impact on stroke occurrence. Our experimental results show that the

Table 1: Experimental results in terms of model accuracy tested through 3-folds cross-validation method when using all 9 features to perform a comparative analysis of the global and personalised modelling approaches.

		Global	Personalised		
Model		SVM (Linear kernal, gamma=1)	KNN (k=165)	WKNN (k=163, threshold=0.5)	FaLK-SVM (k=181, Linear kernal, gamma=-0.5, c=3)
Number of Features		9	9	9	9
Accuracy of Each Class (%)	*Normal*	43.60	43.00	44.40	46.81
	Stroke	49.00	52.40	51.20	56.57
Overall Accuracy (%)		46.30	47.70	47.70	**51.69**

Table 2: Experimental results in terms of model accuracy tested through 3-folds cross-validation method when using 6 features to perform a comparative analysis of the global and personalised modelling approaches.

		Global	Personalised		
Model		SVM (Linear kernal, gamma=1)	KNN (k=165)	WKNN (k=163, threshold=0.5)	FaLK-SVM (k=181, Linear kernal, gamma=-0.5, c=3)
Number of Features		6	6	6	6
Accuracy of Each Class (%)	*Normal*	60.00	63.40	61.00	69.90
	Stroke	69.20	68.00	70.60	70.64
Overall Accuracy (%)		64.60	65.70	65.80	**70.27**

personalised modelling approach outperforms the global modelling approach. Especially, the classification accuracy is dramatically improved when using 6 features as compared using 9 features.

However, this study only looked at the Auckland region and selected 500 samples out of 2850 samples as a preliminary study. Therefore, this work will be further extended to explore all samples in the Auckland region and also other five regions. Furthermore, in the future study, I will try to deal with the categorical data in the classification problems. In addition, from the experimental results, we find out k-nearest neighbour could be an important factor to improve the accuracy, in this study, we manually selected neighbours and model parameters. In the future, we will intend to develop new methods for personalised modelling in order to improve the robustness and generalisability of feature selection, neighbourhood selection, model and its parameter selection for classification, diagnostic and prognostic problems. For instance, the evolutionary algorithm (EA) might be integrated with the personalised modelling approach for solving optimisation problems. EA is inspired by biological evolution, such as crossover, mutation, recombination, and selection to evolve the individuals (candidate solutions) based on the principle of ”fitness to survival”.

References

1. Aizerman, E.M., Braverman, L.R.: Theoretical foundations of the potential function method in pattern recognition learning. Automat Remote Control 25, 821–837 (1964)
2. Baek, O., Gaffney, T., Joshi, K., Robson, B., Rosen, D., Stahlbaum, C., Taylor, R., Vortman, P.: Personalised healthcare 2010: Are you ready for information-based medicine? http://www-935.ibm.com/services/in/igs/pdf/g510-3565-personalized-healthcare-2010.pdf (n.d.)
3. Beygelzimer, A., Kakade, A., Langford, J.: Cover trees for nearest neighbour. In: 23rd International Conference on Machine Learning. Pittsburgh (2006)
4. Chang, C.C., Lin, C.J. LIBSVM: A library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm (2001)
5. Fix, E., Hodges, J.L.: Discriminatory analysis: Nonparametric discrimination: Consistency properties. Randolph Field, Texas (1951)
6. Ginsburg, G.S., McCarthy, J.J. : Personalized medicine: revolutionizing drug discovery and patient care. Trends in Biotechnology 19(12), 491–496 (2001)
7. Iakovidis, I.: Towards sustainable and personalised healthcare, http://ec.europa.eu/information_society/events/phs_2007/docs/slides/phs2007-iakovidis-ch5-1a.pdf (2007)
8. Kasabov, N.: Evolving connectionist systems: the knowledge engineering approach. Springer, London (2007)
9. Khosla, A., Cao, Y., Lin, C.Y., Chiu, H.K., Hu, J.L., Lee, H.: An itegrated machine learning approach to stroke prediction. In: 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, Washington (2010)
10. Lankhorst, M.M., van Kranenburg, H., Salden, A., Peddemors, A.J.H.: Enabling technology for personalizing mobile services. In: 35th Annual Hawaii International Conference on System Sciences. Hawaii (2002)
11. Lesko, L.J.: Personalized medicine: Elusive dream or imminent reality? Clinical Pharmacology & Therapeutics 81, 807–816 (2007)
12. Mukamal, K.J., Wellenius, G.A., Suh, H.H., Mittleman, M.A.: Weather and air pollution as triggers of severe headaches. Neurology 72, 922–927 (2009)
13. Segata, N., Blanzieri, E.: Fast and scalable local kernel machines. Journal of Machine Learning Research 11(2010), 1883–1926 (2010)
14. Vapnik, V.: Statistical learning theory. Wiley-Interscience, NY (1998)
15. Vapnik, V., Lerner, A.: Pattern recognition using generalized portrait method. Automation and Remote Control 24, 774–780 (1963)