Personalized Information Services for Quality of Life: The Case of Airborne Pollen Induced Symptoms

Dimitris Voukantsis[1], Kostas Karatzas[1], Siegfried Jaeger[2] and Uwe Berger[2]

[1]Informatics Systems & Applications Group, Dept. of Mechanical Engineering, Aristotle University, P.O. Box 483, 54124, Thessaloniki, Greece
[2]Medizinische Universität Wien, Universitätsklinik für Hals-, Nasen- und Ohrenkrankheiten, Währinger Gürtel 18-20, 1090 Wien, Austria

Abstract. Allergies due to airborne pollen affect approximately 15-20% of European citizens; therefore, the provision of health related services concerning pollen-induced symptoms can improve the overall quality of life. In this paper, we demonstrate the development of personalized quality of life services by adopting a data-driven approach. The data we use consist of allergic symptoms reported by citizens as well as detailed pollen concentrations of the most allergenic taxa. We apply computational intelligence methods in order to develop models that associate pollen concentration levels with allergic symptoms on a personal level. The results for the case of Austria, show that this approach can result to accurate and reliable models; we report a correlation coefficient up to r=0.70 (average of 102 citizens). We conclude that some of these models could serve as the basis for personalized health services.

Keywords: Allergy, Computational Intelligence, Personalized Health Services.

1 Introduction

Allergy due to airborne pollen is a reaction of the human immune system to certain allergens carried by the pollen grains. It is estimated that approximately 15-20% of the European citizens suffer from pollen-related allergies [1]; additionally, there are well-established, yet not fully understood, associations with certain respiratory diseases, e.g., asthma [2]. During the last years, there has been an increasing trend in pollen-induced allergic symptoms, while the World Allergy Organization [3], reports an increase in severity of the symptoms.

In this context, there is an emerging need for quality of life information services that could assist sensitized citizens in making decisions concerning their daily life. This issue has already been partially addressed by existing information systems that provide information of pollen concentration levels of allergenic taxa [4]. However, the severity of allergic symptoms is strongly dependent on the citizen under consideration. Therefore, a more personalized approach that addresses symptoms rather than pollen concentration levels seems more appropriate.

The design and development of personalized information services addressing citizen-specific symptoms is a challenging task due to the lack of a generic (universal) mechanism responsible for the triggering of allergic symptoms. Moreover, airborne pollen concentration levels have not been regulated as in the case of chemical air pollutants [5], thus there is no regulatory or management framework, based on common limit values and threshold levels to serve as a reference for any health related decision. In order to avoid these difficulties, we adopt a data-driven modeling approach by utilizing the database of pollen-induced symptoms reported by citizens themselves through the Pollen Diary system [6]. Additionally, we use detailed pollen concentration levels, monitored in several sampling sites, for the area of interest. We demonstrate the development of such models for the case of Austria, using data collected during the year 2010. The resulting personalized models could serve as the main modeling tool of existing or future information systems, providing citizen-specific warnings concerning the occurrence and severity of airborne pollen-induced symptoms.

2 Materials and Methods

2.1 Data Presentation

The modeling approach adopted in this study has been based in two distinct databases. i) Pollen concentration data of 65 distinct taxa sampled at several monitoring sites across Austria during 2010, and ii) Pollen-induced symptoms reported by users of the Pollen Diary system. The total number of users was 716, with an average of 57 records per user. The symptoms were reported in detail (eye itching, eye watering, nose, etc.); however, in this case we have identified the “overall symptoms” as the main parameter of interest. The latter one ranges from 0 (no symptoms) up to 25 (strong symptoms in eyes, nose and lungs).

Table 1. Variables included in the preprocessed data set.

Temporal Variables	Pollen Concentration	Symptoms (target)
Day of Year	Acer, Alnus, Ambrosia, Betula, Carpinus, Corylus, Cupress, Fraxinus, Juglans, Pinus, Plantago, Platanus, Poaceae, Populus, Quercus, Rumex, Salix, Sambucus, Tilia, Urtica	Overall Symptoms

The data were preprocessed excluding taxa with more than 10% missing values, resulting to a final of 20 taxa (Table 1). Furthermore, users with less than 100 records were not included in this study, resulting in a total of 102 users. The final preprocessing step was to normalize the data using variance scaling. The latter one results to variables with average value $\mu=0$ and standard deviation $\sigma=1$. Fig.1 presents the overall symptoms (average of all citizens) and average pollen concentrations (average of all taxa) as a function of time. It is evident that during spring (days 80-130) peak values of overall symptoms and pollen concentrations coincide.

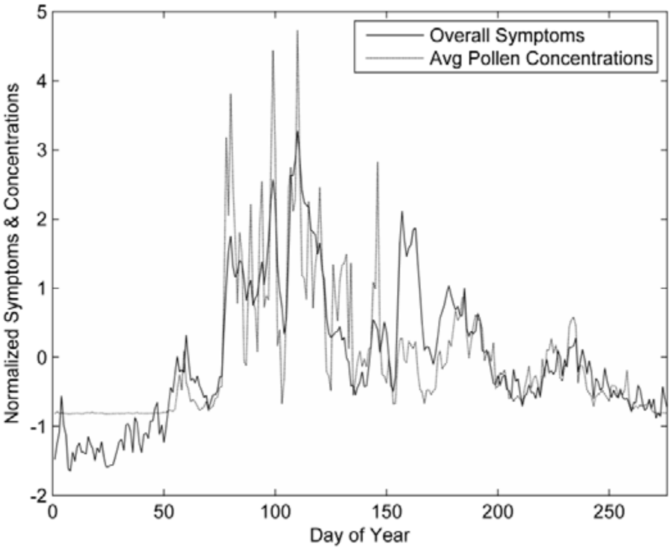


Fig. 1. Overall symptoms (average of all citizens, solid line) and pollen concentration (average of all taxa, dash-dot line). Both variables are normalized, using variance scaling (μ=0, σ=1).

2.2 Computational Intelligence Methods

In order to model the inter-relations between pollen concentrations and citizen-specific symptoms, we have applied a series of Computational Intelligence (CI) methods. The task under consideration has been addressed as a regression problem (the target was the numerical value of "overall symptoms"). Therefore the proper versions of the following CI methods were considered:

Multi-Layer Perceptron (MLP). MLP [7] models are probably the most popular artificial neural network models, capable of modeling complex and highly non-linear processes. In this case, we have evaluated network architectures consisting of one hidden layer, setting the number of neurons equal to *N=(inputs+outputs)/2* [7]. Where *inputs* is the number of input variables of the model and *outputs* the number of output variables; in this case, outputs=1. Furthermore, we have use the Levenberg-Marquardt backpropagation algorithm during the training of the models, with a maximum number of epochs at 500.

Support Vector Regression (SVR). SVR is a modification of the original algorithm introduced by Vapnik [8] capable of tackling regression problems. In this case we have chosen the SVR kernel to be the Radial Basis Function (RBF). Furthermore, we optimized the regularization parameter (*C*) and the ε-intensive zone of the model, as well as the γ parameter of the kernel. The SVR models have been developed using the Spider Toolbox for Matlab [9].

Least Squares Support Vector Regression (LS-SVR). LS-SVR is a modification of support vector machines [10]. In this case, we have used the RBF kernel and optimized the regularization (*C*) and kernel parameter (γ) using leave-one-out cross-validation. The LS-SVR models were developed using the LS-SVMlab toolbox for Matlab [11].

k-Nearest Neighbors (kNN). kNN is non-parametric CI algorithm that can be used to tackle regression problems. In this case, we have set the number of nearest neighbors to be k=5 and we applied the 1/*r* (*r* is the distance to the neighbor) weighting scheme.

Multiple Linear Regression (MLR). MLR is a traditional statistical method that provided the reference frame for comparison of the other CI methods.

2.3 Feature Selection

The selection of the appropriate features as input parameters of the models is an important modeling step as irrelevant or noisy features may result to poor performing models. Mutual Information (MI) as well as other information related criteria have been used in the past in order to identify relevant input variables. In this case, we have used MI following the approach presented in [12].

2.4 Training and Evaluation

Due to the limited number of available instances per citizen, the training of the models was based on the 10-fold cross validation, in order to utilize all instances. The Root Mean Square Error (RMSE), the Correlation Coefficient (r), as well as the Index of Agreement (d) have been used to validate the models. The latter index is defined by the following formula:

$$d = 1 - \frac{\sum_i |p_i - a_i|^2}{\sum_i \left(|p_i - \overline{a}| + |a_i - \overline{a}| \right)^2} \qquad \textbf{(1)}$$

where p_i refers to predicted values and a_i to observed ones, whereas with $\overline{p}$ and $\overline{a}$ are denoted the average of the predicted and observed values, respectively. The index of agreement ranges from 0.0 (theoretical minimum) to 1.0 (perfect agreement between observed and predicted values).

3 Results and Discussion

Table 2 summarizes the performance of the resulting 102 citizen-specific models. The results are presented as averages (and standard deviations) of the statistical indices presented in section 2.4. It is clear that the performance differs significantly depending on the algorithm used to develop the model. The non-parametric kNN models as well as the LS-SVR models indicate superior performance compared to the SVR and MLP models. This may be attributed to the limited number of training examples in the data set (at least for the MLP models). Furthermore, all CI methods perform on average better than the reference (MLR) method. This indicates that CI algorithms are more suitable to model the pollen-induced symptoms.

The results presented in Table 2 indicate that modeling pollen-induced symptoms on citizen level can result to models of acceptable performance. In contrast, a global approach, i.e., using the same data and algorithms to build a model for all citizens, does not result to acceptable performances (d<0.4, r<0.25, rmse>4).

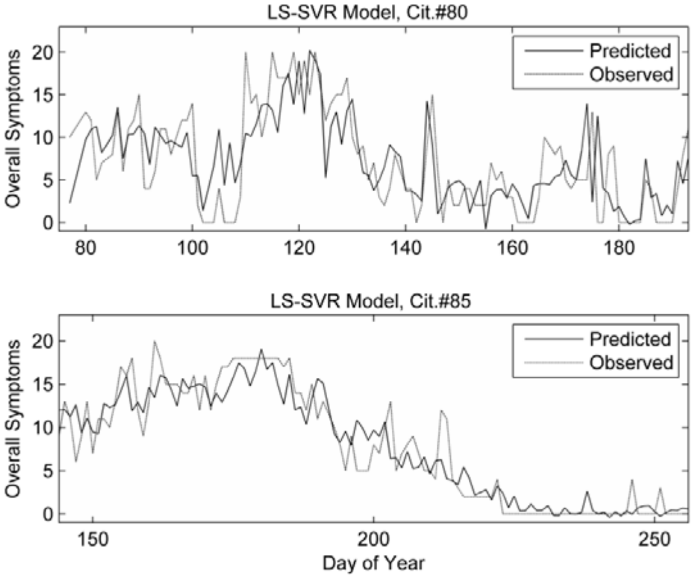
Table 2. Performance (average and standard deviation) of 102 citizen-specific models. The unit of RMSE is the number of symptoms.

Model	d	r	RMSE
MLR	0.74 ± 0.14	0.61 ± 0.20	2.26 ± 1.05
MLP	0.76 ± 0.15	0.63 ± 0.22	2.17 ± 0.93
kNN	0.80 ± 0.11	0.67 ± 0.17	2.07 ± 0.86
SVR	0.74 ± 0.14	0.64 ± 0.19	2.11 ± 0.88
LS-SVR	0.79 ± 0.14	0.70 ± 0.18	1.92 ± 0.80

The prevalence of allergic symptoms in humans is a highly complex process that dependants on several factors such as pollen concentrations, meteorological and chemical (i.e. air quality) weather conditions, citizen habits (outdoor activity, travelling, medication). Therefore, in some cases the available data cannot result to accurate models. This can be demonstrated in more detail for the case of two specific citizens (Citizens with ID number 80 and 85, hereafter denoted as Cit.#80 and Cit.#85), with similar data records and maximum overall symptoms. Table 3 presents the performance of the two citizen-specific models (Cit.#80, Cit.#85). The results show that the allergic symptoms indicated by Cit.#85 can be successfully modeled; however, this is not the case for Cit.#80. Fig.3 presents the overall symptoms for both citizens, as a function of time. It is evident that the pattern of symptoms for both citizens is different.

Table 3. Typical detailed performance (10-fold cross validation) of models for two citizens.

ID	Model	MLR	MLP	kNN	SVR	LS-SVR
Cit.#80	***d***	0.67 ± 0.16	0.73 ± 0.16	0.74 ± 0.15	0.71 ± 0.13	0.79 ± 0.12
	r	0.54 ± 0.23	0.65 ± 0.26	0.59 ± 0.28	0.61 ± 0.19	0.68 ± 0.21
	rmse	4.51 ± 0.95	4.16 ± 1.33	4.44 ± 1.18	4.48 ± 1.08	3.83 ± 1.20
Cit.#85	***d***	0.92 ± 0.09	0.95 ± 0.06	0.97 ± 0.02	0.94 ± 0.02	0.95 ± 0.03
	r	0.88 ± 0.15	0.93 ± 0.06	0.94 ± 0.04	0.92 ± 0.04	0.93 ± 0.05
	rmse	2.87 ± 0.90	2.30 ± 1.02	2.08 ± 0.78	2.70 ± 0.35	2.34 ± 0.67



LS-SVR Model, Cit.#80

LS-SVR Model, Cit.#85

Fig. 2. Overall symptoms (solid line: predicted, dash-dot line: observed) for Cit.#80 and Cit.#85

4 Conclusions and Future Work

In this paper, we have demonstrated the use of CI methods in order to develop models capable of estimating pollen induced symptoms on a personal level. The results show that the data collected by the Pollen Dairy system can provide a solid foundation to build accurate citizen-specific models in most of the cases. The latter ones can serve as the basic modeling tool to develop personalized health services.

Future work in this field includes i) the use more databases and more advanced CI methods in order to improve the performance of the personalized models and ii) the use of the Pollen Diary database in order to categorize citizens into several sensitized groups and identify common characteristics. These goals will advance the accuracy and interpretability of the models, therefore providing a more solid background for the development of personalized health services.

Acknowledgments. The authors greatly acknowledge the contribution of all colleagues involved in the pollen concentration data collection. Manfred Bobek, Sigmar Bortenschlager, Inez Bortenschlager, Uschi Brosch, Edith Bucher, Bernard Clot, Pramod Harvey, Veronika Kofler, Andreja Kofol-Seliger, Herta Koll, Sabine Kottik, Margit Langanger, Rudolf Litschauer, Karol Micieta, Anna Paldy, Ondrej Rybnicek, Hanna Schantl, Jutta Schmidt, Roland Schmidt, Ingrid Weichenmeier, Helmut Zwander.

References

1. Huynen, M., Menne, B., Behrendt, H., Bertollini, R., Bonini, S., Brandao R., et al.: Phenology and Human Health: Allergic Disorders. In: Report of a WHO meeting, Rome, Italy (2003)
2. Taylor, E.P., Jacobson, W.K., House, M.J., Glovsky, M.M.:. Links between Pollen, Atopy and the Asthma Epidemic. International Archives of Allergy and Immunology 144, 162--70 (2007)
3. World Allergy Organization, http://www.worldallergy.org/
4. Europe: Polleninfo.org, http://www.polleninfo.org/
5. CAFE - Derective 2008/6/EC of the European Parliament and of the Council.
6. Pollen diary, https://www.pollendiary.com/Phd/
7. Haykin, S.: NeuralNetworks: A Comprehensive Foundation. Prentice Hall, Upper Saddle River (1994)
8. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag, New York (1995)
9. Spider Toolbox for Matlab, http://people.kyb.tuebingen.mpg.de/spider/index.html
10. Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J.: Least Squares Support Vector Machines. World Scientific, Singapore (2002)
11. LS-SVMlab, http://www.esat.kuleuven.be/sista/lssvmlab/
12. Peng. H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27,8, pp. 1226--1238 (2005)