Sleep quality analysis

Arturo Laflor

2017-04-27

Contents

Chapter 1

Prerequisites

Chapter 2

Introduction

Chapter 3

Data adquisition

Chapter 4

Data pre-process

Chapter 5

Feature selection

This section describes the process that we perform to reduce the dimension of the sleep hygiene data set that contains the features to model the quality of sleep for respondents of the survey described in chapter 3. There exist two ways to addressed dimensionality reduction, feature extraction and feature selection. Feature extraction, consists in generate a new and small feature space. The application of a technique of feature extraction produce new features based in original ones. The new dataset is not understandable in terms of the original dataset, rather, it is an abstraction of this and its visualization have no practical meaning. On the other hand, feature selection choose a small subset of the relevant features from the original dataset according to certain relevance evaluation criterion, which usually leads to better learning performance, lower computational cost, and better model interpretability (Tang et al., 2014).

For the purposes of this study, the technique of selection of characteristics is the most appropriate. Our interest in reducing dimensionality is not related to the decrease in computational cost, rather, the purpose is to decrease the number of predictive variables due to the high cost of design and infrastructure that means capturing 21 different signals through sensors. If it is possible to characterize a high percentage of the phenomenon, through a reduced number of factors of sleep hygiene, the design of the system will be more feasible and less expensive.

The model accuracy for prediction of the sleep quality with the subset of features must be better than the training model using the total of sleep hygiene features.

5.1 Feature selection models

In 1996, (Liu and Motoda, 1998) proposes two models to achieve the reduction of features, that have been used as basis of diverse algorithms still in force. The filter model (see Figure 5.1) that uses as criterion of feature selection, some attributes concerning only to the data domain. Especifiacally in this model, Liu et. al. proposes that it is posible to analyze and make decisions over irrelevance or relevance of features based in measure information gain, dependence, distance and consistency.

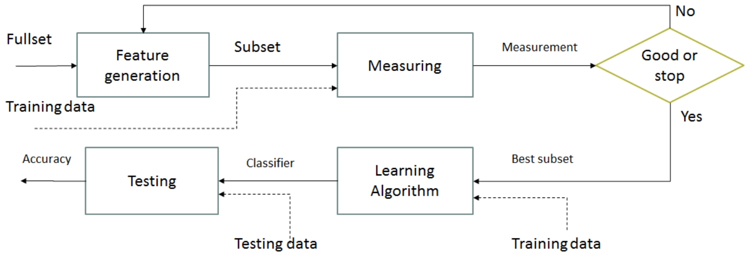

Figure 5.1: Filter model proposed by Liu et. al.

The second model showed in Fig. 5.2 proposed is the wrapper model that uses the accuracy of prediction as selection criterion, it means that this techniques are committed with a particular classifier in this stage of the learning process.

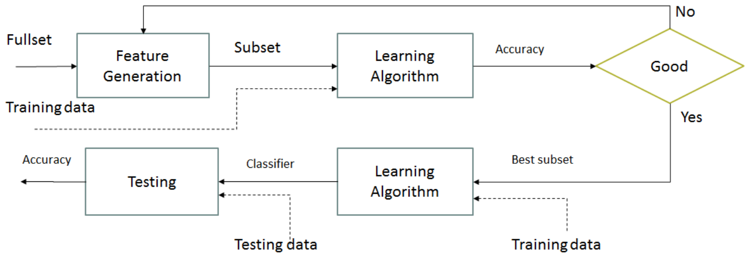

Figure 5.2: Wrapper model proposed by Liu et. al.

Both models have advanteges and disadvantages, techniques based in filter model, performs better than others based in wrapper model, however, researchers have no idea over prediction accuracy during the feature selection process. Some practitioners don't preffer to use these techniques because if accuracy prediction is not achieved in the proposed level, the first steep can be regarded as a waste of time. On the other hand, some researchers argued that select features based in determinated classifier, reduces the possibility to use other classifier to generate the prediction model, in this sense, the classifier to generate the final model should be choosed at the begining, and it is not convenient for all problems. In these order of thinks, (Kelleher et al., 2015) comment that wrapper models are more computationally expensive than filters models and that the argument of they are uncertain models respect to the accuracy, is not at all valid since filters model often generate models with good accuracy.

Additionaly, (Liu and Motoda, 1998) highlight *Search*, *Scheme* and *measure* as three important concepts that help to decide what technique is the most appropriate for an specific problem of dimentionality reduction by feature selection (see Fig. 5.3). Search refers to the activity of choose features in non deterministic, heuristic or complete form, Scheme must be determine if the search will be forward, backward or in random mode, and, measure has to do with tree ways to establish the threshold for stopping the feature search, the criterion used are accuracy, consistency, and, classic criterion involving distance, information gain and dependence.

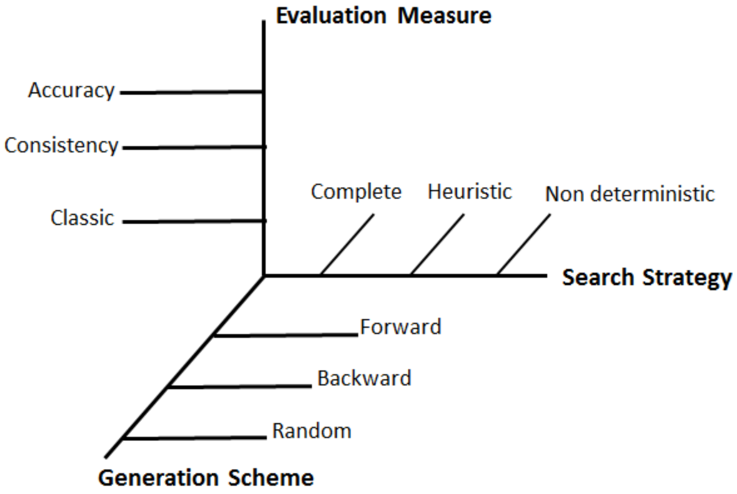

Figure 5.3: Main dimensions in feature selection, Liu et. al.

A third type of model has been proposed in last years, these models are called **embedded models**, since they allow practitoners select features while the prediction model is built. Embedded models have the advantage of filters model in terms of low computational cost, and take the advantage of wrapper model, because the prediction accuracy and classification model are involved in the process. (Tang et al., 2014) describe three type of embedded methods as we shows in the Table 5.1.

Table 5.1: Embedded methods as Tang et al. (2014) describes and quoted verbatim in his paper.

Method	Description	Cite
Pruning	Utilizing all features to train a model and then attemp to eliminate some features by setting the corresponding coefficients to 0, while maintaining model performance such as recursive feature elimination using support vector machines (SVM)	Guyon et al. (2002)
Build-in	Mechanism for feature selection as ID3 and C4.5	Quinlan (1986, 1993)
Regularization	Utilices objective functions that minimize fitting errors and in the mean time force the coefficients to be small or ti be exact zero.	Ma and Huang (2008)

These models are representative of the theoretical basis where a lot of algorithms for selection features in last twenty years have been fueled. Likewise four concepts are the most important and have been used for the generation of different feature selection algorithms in last two decades: distance, accuracy, inconsistency and information gain.

- Distance: The main goal to use distance, is to find similarity among instances in a dataset. The Equation proposed by Minkowski (see (**??**)) is a generalization of the distances that are used in MLA. The most common distances are the particular cases where $p = 1$ called Manhatan distance (see Eq. (5.2)) and where $p = 2$, the well known Euclidian distance (see Eq. (5.1)). (All three equations were taken from (**?**))

$$Euclidean(A, B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_1)^2 + \cdots + (a_n - b_n)^2} \tag{5.1}$$

$$Manhattan(a, b) = \sum_{i=1}^{m} abs(a[i] - b[i]) \tag{5.2}$$

$$Minkowski(a,b) = \left(\sum_{i=1}^{m} abs(a[i] - b[i])^{p} \right)^{\frac{1}{p}} \tag{5.3}$$

The implication of use different values of p will be noted in the difference between two values of any feature in the final distance, it is directly proportional to the value of p. It means that large differences between two features in an instance, impact stronger in the final result when p grows.

- Accuracy: Accuracy refers to the successes that a model had to predict each instance of a dataset, it is opposed to the miscalssification error as (Kelleher et al., 2015) defines in (5.4) and (5.5) equations.

$$misclassification\ rate = \frac{(FP + FN)}{(TP + TN + FP + FN)} \tag{5.4}$$

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{5.5}$$

Chapter 6

Methods

Chapter 7

Applications

Chapter 8

Placeholder

Chapter 9

Final Words

Bibliography

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

Kelleher, J. D., Namee, B. M., and D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: algorithms, worked examples, and case studies.* Number 1. The MIT Press, London.

Liu, H. and Motoda, H. (1998). *Feature selection for knowledge discovery and data mining.*

Ma, S. and Huang, J. (2008). Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, 9(5):392–403.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Quinlan, J. R. (1993). *C4.5: programs for machine learning.* Morgan Kaufmann.

Tang, J., Alelyani, S., and Liu, H. (2014). Feature Selection for Classification: A Review. *Data Classification: Algorithms and Applications*, pages 37–64.