Towards Conceptual Predictive Modeling for Big Data Framework

Jeong-Sig Kim¹, Eung-Sung Kim² and Jin-Hong Kim³

¹,² Department of Computer and Mobile Convergence, Gyeonggi College of Science and Technology, 269, Gyeonggigwagi-dearo, Siheung-si, Gyeonggi-do, Korea
³ College of Information and Communication Engineering, Sungkyunkwan University, 2066, Seobu-ro, Jangan-gu, Suwon-si, Korea
¹,²{arius70, imagecap}@gtec.ac.kr, ³jinhkm@skku.edu

Abstract

Predictive modeling is the process of creating a statistical model from data with the purpose of predicting future behavior. In recent years, the amount of available data has increased exponentially and “Big Data Analysis” is expected to be at the core of most future innovations. Due to the rapid development in the field of data analysis, there is still a lack of consensus on how one should approach predictive modeling problems in general. Another innovation in the field of predictive modeling is the use of data analysis competitions for model selection. This competitive approach is interesting and seems fruitful, but one could ask if the framework provided by for example Gane Project based on big data framework gives a trustworthy resemblance of real-world predictive modeling problems. In this thesis, we will state and test a set of hypotheses about predicative modeling, both in general and in the scope of data analysis competitions for model selection. This competitive approach is interesting and seems fruitful, but one could ask if the framework provided by for example Gane Project based on big data framework gives a trustworthy resemblance of real-world predictive modeling problems. In this thesis, we will state and test a set of hypotheses about predicative modeling, both in general and in the scope of data analysis competitions for model selection. We will then describe a conceptual big data framework for approaching predictive modeling problems. To test the validity and usefulness of this framework, we will participate in a series of predictive modeling competitions on the platform provided by Gane, and describe our approach to these competitions.

Keywords: Conceptual Predictive Modeling, Gane system, Big data Framework

1. Introduction

Prediction is a central problem in machine learning that involves inducing a model from a set of training instances that is then applied to future instances to predict a target variable of interest [1]. Several commonly used predictive algorithms, such as logistic regression, neural networks, decision trees, and Bayesian networks, typically induce a single model from a training set of instances, with the intent of applying it to all future instances. We call such a model a population-wide model because it is intended to be applied to an entire population of future instances [2]. A population-wide model is optimized to predict well on average when applied to expect future instances. Recent research in machine learning has shown that inducing models that are specific to the particular features of a given instance can improve predictive performances. We call such a model an instance-specific model since it is constructed specifically for a particular instance (case) [3-5]. The structure and parameters of an instance-specific model are specialized to the particular features of an instance, so that it is optimized to predict especially well for that instance. The goal of inducing an instance-specific model is to obtain optimal prediction for the instance at hand. This is in contrast to the induction of a population-wide model where the goal is to obtain optimal predictive performance on average on all future instances. There are several possible approaches for learning.
predictive models that are relevant to a single instance [6]. One approach is to learn a model from a subset of the training data set that consists of instances that are similar in some way to the instance at hand. Another approach is to learn a model from a subset of variables that are pertinent in some fashion to the instance at hand. A third approach, applicable to model averaging where a set of models is collectively used for prediction, is to identify a set of models that are most relevant to prediction for the instance at hand.

2. Conceptual Modeling

The main problem in modeling educational data is to discover or learn robust and accurate models from small data sets. The most important factor which affects the model accuracy is the model complexity. Too complex models do not generalize to other data sets, while too simple models cannot model the essential features in the data. Model complexity has an important role in robustness, too, but there are several other factors which affect robustness. Each modelling paradigm and model construction method has its own inductive bias – a set of conditions, which guarantee a robust model. These conditions should be checked when we select the modelling paradigm, model structure, and model parameters. In the following, we will first define the basic modelling terminology. Then we analyze the main factors – model complexity, inductive bias and robustness – which affect the model selection. We formulate principles for selecting an appropriate modelling paradigm and a model in the educational domain. Finally, we summarize the main techniques for model validation. In data-driven modelling the models are learnt from a finite set of samples called a training set. Because the training set has only a limited size, we seldom find the true model, which would describe all possible data points defined in the relational schema. Thus, the learnt model is only an approximation [7-8]. We show the 4-Definition as bellows:

Definition 1 (True and approximated model) Let R be a set of attributes and r according to R a relation. Let X = {X₁, ..., Xₖ} ⊆ R be a set of attributes in R and Y ∈ R the target attribute. Then the true model is a function F : Dom(X₁) × ... × Dom(Xₖ) → Dom(Y) such that t[Y] = F(t[X]) for any tuple t defined in R. The approximated model is a function approximation M : Dom(X₁) × ... × Dom(Xₖ) → Dom(Y) such that t[Y] ≈ M(t[X]) for all tuples t ∈ r.

Definition 2 (Partial model) Let R, r ∈ R, X ⊆ R and Y ∈ R be as before, and r 0 ( r. Model M is partial, if M|r 0 : Dom(X₁) × ... × Dom(Xₖ) → Dom(Y) such that t[Y] ≈ M(t[X]) for all tuples t ∈ r.

Definition 3 (Training error and true error) Let d(M(X), Y) be an error between the predicted value M(X) and the real value of Y. Then the training error of model M in relation r, |r| = n.

Definition 4 (Optimal model) Let M be a model, r a relation, and score(M, r) a function, which evaluates the goodness of model M in relation r. We say that the model is optimal, if for all other models M₀ ̸= M, score(M₀, r) ≤ score(M, r).

3. Big Data

Data is the main subject of knowledge discovery. The selection of attributes, their types and domains have a strong influence on the model accuracy. In this paper, we demonstrate that with careful data preprocessing we can model even small educational data sets accurately. In the following, we will define the main concepts concerning data, analyze typical characteristics of educational data, and propose appropriate preprocessing techniques [9-11].
3.1. Gane Data

In the Gane-project, the first problem was to gather data from several sources and determine the best attributes for our modeling purposes, especially for detecting drop-outs and potential failures. Data types classify the attributes according to their domains. The basic division is to numerical (quantitative) and categorical (qualitative) data types. Numerical data has meaning as numbers and we can measure order, distance and equality of two numerical variables. In contrast, categorical data has no meaning as numbers, and generally we can measure only equality between two categorical variables. The general classification is presented in Figure 1. Numeric data can be further classified as discrete or continuous. Discrete data can get only countably many values, and if the domain is pictured on the number line, it consists only of isolated points. Discrete numbers are not necessarily integers, but also decimal numbers with a predefined precision. For example, in the Gane data, the total course points could get any values in the set \{0.00, 0.25, 0.50, ..., 33.50, 33.75, 34.00\}. All exercise task points were also given in 0.25 point precision. On the other hand, continuous data can in principle get any value in the given interval [12].

![Figure 1. The Roles of Pattern Recognition and Information Retrieval for Knowledge Discovery Process](image)

4. A Framework for Predictive Modeling

In this paper, we will outline a conceptual framework for approaching predictive modeling competitions – from the initial data exploration and preprocessing to final model selection and combination. For each part of the framework, we will try to describe common methods, tricks and things to consider [13].

4.1. Exploratory Analysis of Gane data

The first step when doing data analysis, should always be to explore and understand the data. Sometimes the data is rather simple and manageable, but sometimes data exploration can be very time consuming. There is no perfect recipe for how to perform an adequate exploratory analysis, it is a matter of experience and hard work. Still, there are some things which are often a good idea to consider when exploring the data. **Outlier Detection (OD).** An important reason for exploring the data is to see if there is “wrong” data in the provided dataset. This can be measurements which violates some kind of assumption. In our recent work with data analysis competitions, some of the things which have come up in this category are: Probabilities not summing to one, products which are sent for repair before
even being sold and hotel prices of a million dollars per night. The best way to find outliers is to either inspect the minimum value, the maximum value and the percentiles of the data numerically, or more commonly to make plots of the different variables [14]. **Plotting the Gane data (PGD).** One of the best ways to explore data is to visualize it. If one has thousands or maybe even millions of data samples, exploring them manually is not a real possibility. Using the right kinds of visualizations can give the insight into the data which is needed. There is a vast amount of literature and research on how to make good data visualizations, so we will simply mention some of the most common and trivial ones here. The most common and simple plot is the scatter plot as shown in Figure 2. It shows the relation between two numerical variables using points.

![Figure 2. Two Scatter Plots Showing the Connection between the Sepal Width and the Petal Length](image)

Using histograms as shown in Figure 3, one can visualize a single variable. Values close to each other are combined in bins [15], and the height of the bins indicates the number of samples binned together.

![Figure 3. Two Histograms Showing the Distribution of Sepal Length](image)

### 4.2. Feature Engineering

Predictive analysis is all about having the correct features. In some situations, there are no directly useful features and in other situations there is a vast amount of irrelevant or correlated features. In this paper, we will try to describe some of the common techniques for handling feature-related issues in predictive analysis. We will split the section into two parts, one about reducing the number of features (by removing the least useful features), and one about generating and extracting new useful features [16]. **Feature Reduction (FR).** A simple and yet very effective way to reduce the number of features, is to manually remove some of them based on intuition about their importance. As an example, say you want to predict the value in dollars for which a product is sold on an online auction (like eBay). On the other hand, the second that the product was posted does most likely not hold any information about the sale price. Removing non-useful features will make the statistical models simpler, more robust and less prone to
overfitting. The basic idea behind using random forests to measure variable importance is the following: A random forest is fitted to the data, and the out-of-bag error measure is calculated. To measure the importance of feature \( j \), we permute the values in feature \( j \) and compute the out-of-bag error again on this permuted dataset. If the out-of-bag estimate increases a lot by permuting feature \( j \), then that feature must have been important for the predictions. Consequently we rank the features by the difference in out-of-bag error we get by permuting them, normalized in an appropriate way.

5. Algorithmic Evaluation

A meaningful methodology for algorithmic evaluation is needed for at least two reasons: to demonstrate the capabilities of an algorithm in a particular application and thus estimate its effectiveness; and to provide a systematic method for evaluating (perhaps incremental) changes to algorithms. There has been much good work in the past few decades in the development of algorithms for the extraction of various types of information from images. This work has generally concentrated on the assumptions that must be made regarding the data and the numerical form of possible solutions. Often quite strong assumptions as to the characteristics of the data are imposed for reasons of mathematical tractability. Much less work has been published on the systematic evaluation of algorithms in terms of relaxing these assumptions or for the purposes described above. This may in part be due to the fact that a rigorous evaluation is a large amount of extra work and has often not been perceived as publishable in the same way as a novel piece of mathematics or the demonstration of a new application. Systematic evaluation may require a completely different approach to algorithm design and testing. Having developed an algorithm over a period of several years, the complexity of the resulting algorithm may be such that it is impossible to evaluate the algorithm in any way other than treating it as a black box. As complexity increases, it becomes progressively harder for such a black box evaluation to provide accurate performance predictions to a potential user. This is caused by the increased number of possible discontinuities due to special cases likely to be present. Furthermore, if the user intends to use the algorithm as a part of a larger automatic system, the quality of the output data needs to be suitable for the next stage in the system. Indeed, it may be argued that if the system is to be used in an application with any social or economic value, a simple algorithm with predictable performance may be better than a complex algorithm with better mean but less predictable performance. In short, algorithms need to deliver not only the answer but accuracy and confidence estimates if the data are to be used reliably in a system. Any algorithm that can only be evaluated as a black box may never produce such output. This suggests that the way to develop good algorithms is to perform algorithm evaluation hand-in-hand with increasing complexity, adding new stages to an algorithm only when the effects of the change can be adequately modeled. This is a more rigorous, though perhaps slower, approach to algorithmic development than is generally followed.

6. Related Works

In general we can define big data comprehensively by the three Vs (Volume, Variety and Velocity) which are commonly used to characterize different aspects of big data [17]. **Volume.** The volume in big data refers to the amount of data that is larger than the capacity of conventional relational database infrastructures. The volume introduces the first challenge of conventional IT structure. We have two options for processing these large volumes of data: one option is the massively parallel processing architectures, and the other is Apache Hadoop based distributed batch processing solutions. The data warehousing approaches need a predetermined schema which is suitable for regular and slowly evolving database. However, Apache Hadoop has no constrain over the structure of the data it can process which make Apache Hadoop suitable for many applications
when we deal with semi or unstructured data. However, the choice is often influence by the other Vs where velocity comes into play. *Velocity*. Real time analytics are becoming increasingly popular. There is a huge demand to analyze fast-moving data which is often referring as "streaming data" or "complex event processing" in the industry. There are two main reasons to consider streaming processing. One provide when the input data are too fast to store in database and some level of analysis is needed when the data streams in. The other presents when we want to get immediate response to the data. *Variety*. One of the other aspects of big data is the demand of analysis of semi-structured and unstructured data. Data could be in text from social networks, images or raw feeds from different sensor sources. The challenge here is to extract ordered meaning for either humans or as structured input for other applications. In relational database we need predefined schemas which result in discarding a lot of data that cannot be captured by a predefined schema. However, the underlying principle of big data is keeping everything as long as possible because the useful signals might be hidden in the bits we throw away. The NoSQL databases fulfill the need for flexibility to store and query semi-structure and unstructured data. They provide enough structure to organize data without constraining fixed schema. Moreover, we can use document store, key value store, column oriented databases depending on our application [17].

6.1. Complexity

Complexity of the data source can be assessed in four different aspects: data structure, data format, data itself and hierarchies used in the data [19]. The way in which the data is received, read, validated, processed and stored by this depends on the characteristics of the data. Complexity of the data structure means that there can exist various relations between data, including complex keys in tables, that makes it difficult to integrate various data tables. Depending on the type of the data source it may be related to the number and size of unstructured files that must be integrated to create the unified data source [20-23]. Complexity of the data format, complexity can be measured by checking what kind of data standards were used to store the data. Complexity of the data source can be regarded as complex when there is lack of information on the code lists used in the data or the code lists used in the data are not integrated in one data source. Hierarchical complexity; this reflects the extent of hierarchies and nested structures in the data. Depending on the requirements of statisticians it can be difficult to drill-down to a specific level of the data [24]. For output, complexity is only relevant if the output is in unit record form that may reflect the complexity of the input, or if there is a need to report on how complexity in input data has been dealt with in the previous stages and if it caused any limitations to the outputs.

6.2. Completeness

Completeness is the extent to which metadata are available for a proper understanding and use of data. It refers to the exhaustiveness of the descriptions available for the input data (*i.e.*, covering all the required aspects mentioned in the hyperdimensions Source and Data as well as the level of detail of descriptions. It includes descriptions of objects (populations, units, and events) [25-27], variables and reference times as well as applied procedures for data treatment and quality measures or qualitative assessment of input data quality. Access to the data file record layout can be considered a minimal requirement for use. For complex data files, the usefulness of the data will depend heavily on the type of information available about the file structure, coding and classification variables. In some cases, the absence of relevant information may drastically limit the potential use of the data if this information cannot be deducted or evaluated from the data itself. The completeness and interpretability assessment of the documentation should cover the steps
necessary for the evaluation at the input stage but also for the subsequent stages (throughput and output).

6.3. Usability

The usability of a dataset is the extent to which this will be able to work with and use the data without the employment of specialised resources or place significant burden on existing resources; and the ease with which it can be integrated with existing systems and standards.

One of the features of Big Data is an increasing diversity in data types, structures and formats. A typical it is structured around the receipt and processing of long-standing data sources, most notably survey and census data, and more recently some kinds of administrative data. This will be able to more easily make use of an incoming dataset that is compatible with existing systems and expertise [28-32]. For new, varied sources of data and new methods of processing and analysing those data sources, new systems and infrastructure will need to be developed. While there is a strong incentive for it’s to develop new capabilities, the extent to which this will need to be done for any particular data source is an important consideration in considering the quality of the data source. If new expertise or infrastructure is needed, the question then arises about the extent to which the improved capability will be transferrable to other data sources [33]. Development of new capabilities provides benefits to this but also comes at a cost; this trade-off is something that should be taken into account when considering a data source.

7. Conclusions

In this paper, we investigated competitive predictive modeling, with a main focus on the big data framework. We have to describe more applicable methodologies for predictive modeling, we outlined a conceptual framework. This framework describes the process from data exploration, over data pre-processing to model training and finally selecting the right model. We describe common tricks and techniques for different types of problems and provide some insight into the minds of the best predictive modelers. In addition, as an attempt to describe more applicable methodologies for predictive modelling, we outlined a conceptual framework.

References