Chapter XIII
Text Mining to Define a Validated Model of Hospital Rankings

Patricia Bintzler Cerrito
University of Louisville, USA

ABSTRACT

The purpose of this chapter is to demonstrate how text mining can be used to reduce the number of levels in a categorical variable to then use the variable in a predictive model. The method works particularly well when several levels of the variable have the same identifier so that they can be combined into a text string of variables. The stemming property of the linked words is used to create clusters of these strings. In this chapter, we validate the technique through kernel density estimation, and we compare this technique to other techniques used to reduce the levels of categorical data.

INTRODUCTION

Nominal data of any type usually require some accommodation before it can be used in comparison studies, particularly if there is a large number of levels in the nominal field. At the same time, nominal data frequently consist of nouns only so that such variables are not normally analyzed using text mining. However, if there is also an identifier field such that multiple items of the nominal field can be linked to one identifier code, then it is possible to use text mining to group and rank clusters of levels in the nominal data field. Once the levels are clustered, an outcome variable can be used to rank the clusters.

The process generally involves the following steps:

1. Transpose the data so that the observational unit is the identifier and all nominal values are defined in the observational unit.
2. Tokenize the nominal data so that each nominal value is defined as one token.
3. Concatenate the nominal tokens into a text string such that there is one text string per identifier. Each text string is a collection of tokens; each token represents a noun.
4. Use text mining to cluster the text strings so that each identifier belongs to one cluster.
5. Use other statistical methods to define a natural ranking in the clusters.
6. Use the clusters defined by text mining in other statistical analyses.

In this chapter, we will demonstrate steps 1-6 to show how nominal data related to patient medical conditions can be used to define a patient severity index needed to model the quality of health care providers. We will first define the problems associated with determining health care provider quality using existing methods (Background into Patient Severity Indices) followed by the application of text mining (Solutions Using Text Analysis). The section, Predictive Modeling with Text Clusters, in this chapter will show how the text clusters defined by text mining can be used in other applications that are relevant to health care.

BACKGROUND INTO PATIENT SEVERITY INDICES

Problems with Terminology Definitions

In many cases, the analysis of a set of data depends almost entirely upon data definitions used during data entry. If entities rely on different definitions and comparisons are made in entity performance, then it is possible to “game” the system to receive a favorable ranking. For example, consider the issue of infant mortality. It seems a relatively simple expression, one with a fairly concrete definition. However, is infancy one year or two? Does infancy begin at birth regardless of the gestational age of the infant? The definition of the World Health Organization is any death that occurs from the moment of the birth of a living child through the first year of life. It is the definition used in the United States. However, many European nations do not count infants less than 500 grams, or less than 28 weeks gestational age in the infant mortality numbers (Buitendijk & Nijhuis, 2004; Cardlidge & Stewart, 1995; Gourbin & Masuy-Stroobant, 1995; Graafmans, Richardus, Macfarlane, Rebagliato, Blondel, & Verloove-Vanhorick, 2001; Joseph, Allen, Kramer, Cyr, & Fair, 1999). By eliminating infants at highest risk of death, the numbers will look more favorable compared to nations that do count those infants. By using a more stringent definition, the United States ranks lower compared to European nations that do not count similar deaths.

An even more difficult term is that of “health”. A recent study published in JAMA indicated that men in the United Kingdom are healthier compared to men in the United States (Banks, Marmot, Oldfield, & Smith, 2006). The definition of “health” was determined to be whatever health factors could be examined using self-report studies that are readily available. For example, the study looked at self-reported rates of cancer and concluded that American men have a higher rate of occurrence, and used that occurrence as an indication of poorer health. However, British men have a higher rate of cancer mortality, which suggests that cancers may not be detected for early treatment. In other words, cancer occurrence must be considered in relationship to cancer screening and early detection. Moreover, the study did not differentiate between types of cancer. British men report a higher rate of smoking and obesity compared to American men. These lifestyle factors were discounted in the authors’ conclusions.

In addition, the study reports the results of an HbA1c test as a method for screening for diabetes. This test gives the average blood sugar level over time. However, this test is not the most common screening tool for diabetes, and most men will
Related Content

A Survey of Parallel and Distributed Data Warehouses
[www.igi-global.com/chapter/survey-parallel-distributed-data-warehouses/53075?camid=4v1a](www.igi-global.com/chapter/survey-parallel-distributed-data-warehouses/53075?camid=4v1a)

On the Mining of Cointegrated Econometric Models
[www.igi-global.com/chapter/mining-cointegrated-econometric-models/39641?camid=4v1a](www.igi-global.com/chapter/mining-cointegrated-econometric-models/39641?camid=4v1a)

Computing Join Aggregates Over Private Tables
[www.igi-global.com/article/computing-join-aggregates-over-private/1816?camid=4v1a](www.igi-global.com/article/computing-join-aggregates-over-private/1816?camid=4v1a)

SeqPAM: A Sequence Clustering Algorithm for Web Personalization
[www.igi-global.com/article/seqpam-sequence-clustering-algorithm-web/1777?camid=4v1a](www.igi-global.com/article/seqpam-sequence-clustering-algorithm-web/1777?camid=4v1a)