Predicting the Likelihood of Falls among the Elderly Using Likelihood Basis Pursuit Technique

Kanittha Volrathongchia Ph.D. Khon Kaen University, Thailand
Patricia F. Brennan, RN, Ph.D., FAAN, FACMI University of Wisconsin-Madison
Michael C. Ferris, Ph.D. University of Wisconsin-Madison

ABSTRACT

This study reports on the application of the knowledge discovery in database process to generate models that can predict the likelihood of falls among the elderly who reside in long-term care facilities. This process was applied to data held in the Minimum Data Set, a comprehensive resident assessment instrument being used in all Medicare and Medicaid supported nursing homes in the United States. For this study, we incorporated a new data mining technique, Likelihood Basis Pursuit, into the process. Using this technique, we were able to correctly identify which of the variables in this data set were associated with falls and generate models that could make fall likelihood predictions based upon those variables. Because the model provides probabilities based upon the exact combination of variables present in a particular resident, models constructed using this new data mining technique have the potential to be more useful for assessing fall risk.

INTRODUCTION

Currently, with the advances in information technology, health data have the potential to be collected on a massive scale in a systematic and electronic form. We are able to draw insight from such data for clinical knowledge development. [1] However, due to the massive scale, many relationships may remain hidden because of the limitations in human ability to recognize them. Thus, in order to utilize these data effectively, we also need to develop automated tools that can be used to extract patterns or relationships among the variables in this collected data.

The purpose of this study is to demonstrate and evaluate the application of one such automated tool. In this study, the Knowledge Discovery in Databases (KDD) process is used to generate models to predict the likelihood of falls among the elderly residing in long-term care facilities. KDD is a collection of processes involving many steps, requiring decisions to be made as to how best to accomplish each step. One critical step is the data mining step. Numerous data mining techniques exist that could be used for this step. In this study, we evaluated the relatively new Likelihood Basis Pursuit (LBP) technique.

BACKGROUND

In order to gain useful information and/or knowledge from large health data sets that are beginning to become available, accompanying advances in automated data mining tools must be made. Data mining, a critical step in the KDD process, has been used for analysis, modeling, and prediction in many health-related areas such as bioinformatics, medicine, and nursing. [2,3] The advantage of data mining techniques is in their automated nature that allows many variables to be examined for relationships to other variables including outcome variables. More sophisticated data mining tools will allow us to gain more insight into the knowledge contained in these massive, information-rich databases.

Different data mining techniques are suitable for different situations. In nursing, data mining tools have historically been applied to classification tasks. [3,4] Classification tasks group people into categories, such as yes/no or risk/no-risk categories. However, in health care, simple classification may not adequately represent the complexity involved in human health needs. Thus, it is often more useful to know the likelihood that a particular health problem might occur in the near future.

For example, an elderly person with a history of falls and a visual problem may have a likelihood of fall greater than one who has a history of falls, but normal vision. In such a case, a classification model is likely to classify both of these individuals into the *fall* category. This occurs even though these two individuals have a different likelihood of falling and require different interventions based upon their individual conditions. However, if we were to know the probability that a health problem might occur in the near future for an individual, we would be able to provide the appropriate interventions to that particular client.

Furthermore, some classification data mining techniques (such as Support Vector Machine) provide predictive results without any indication of how the individual variables in the database contributed to the prediction. In order for predictions to inform practice, the predictions must provide some information about the variables used. For example, in the case of falls, nurses can only provide helpful interventions if they know both who is likely to fall and which variables put that person at high risk for falls. Unlike most classification techniques, the LBP data mining technique does provide this information.

For this reason, we chose to employ the LBP data mining technique in this study as part of the KDD process to generate models to predict the likelihood of falls among the elderly residing in long-term care facilities. The LBP technique is a non-parametric penalized likelihood approach, proposed by Zhang et al. [5] as a flexible nonparametric alternative to the parametric approaches for variable selection and model building. Unlike other data mining techniques that look

for a classifier, LBP derives a probability estimate for the outcome given explanatory vectors while automatically selecting important variables. Models constructed using the LBP process determine the probability by maximizing the log-likelihood and minimizing the penalty basis pursuit. This technique was derived from a combination of the Smooth Spine Analysis of Variance (SS ANOVA) and basis pursuit.[5]

A few past studies have made use of models constructed using the LBP data mining technique. For example, the LBP technique was employed to identify the possible risk factors for the progression of diabetic retinopathy. [6] The LBP technique was also employed to identify the possible risk factors for five-year mortality for non-diabetes participants. [5] These studies show that the LBP technique can select important variables effectively and that the results are comprehensible.

This method was chosen, in part, for this study because it had the ability to determine the relationships among a large number of variables in a nonparametric manner, which would overcome the limitations of standard parametric methods. It can also deal with categorical and numerical variables simultaneously, both of which are contained in the data set chosen for this study, the Minimum Data Set (MDS).

The phenomenon of falls was chosen for this study because falls are a common health problem for the elderly and can lead to serious consequences. Epidemiological studies have reported an annual fall incidence in long-term care (LTC) facilities of 1,500 to 3,000 falls per 1,000 residents. [7] The average incidence rate is around 1.5 falls per bed per year. [7] Furthermore, Thapa et al.,[8] reported that 45% to 75% of LTC residents fall annually, twice the rate of community–living older adults. The associated morbidity and mortality greatly impacts quality of life. According to epidemiological studies [7], around 4% of falls result in fractures, whereas other serious injuries such as head trauma, soft-tissue injuries, and severe lacerations occur in about 11% of the cases. Each year, about 1,800 fatal falls occur in nursing homes. [7] Furthermore, falls lead to serious physical functioning, quality of life, and psychological consequences. Loss of function can result from both fracture-related disabilities and self-imposed functional limitations caused by the fear of falling. Given the incidence of falls and the potential for injury imposed by falls, a fall is one of the most common health problems for LTC residents. Moreover, the literature on falling is well established, allowing us to compare our results directly with established knowledge.

Other studies have made use of the KDD process to generate predictive models from existing databases. However, this study expands upon that by using a new data mining technique, LBP, to accomplish this. The model constructed in this study supplied information about both the likelihood of falls and the variables that contributed to the likelihood predictions. Other data mining techniques do not necessarily provide this information. The addition of this information might help health care personnel better understand the risks faced by the elderly in general and the underlying phenomenon of falls in particular.

METHODS

This non-experimental study employs KDD to do secondary data analysis of the Minimum Data Set (MDS) data obtained from LTC facilities in Kansas in 1996, and acquired from the Centers for Health Systems, Research, and Analysis (CHSRA), University of Wisconsin-Madison.

The MDS is a comprehensive resident assessment instrument (RAI) that measures functional status, mental health status, and behavioral status of the residents residing in LTC facilities to identify chronic care patient needs and formalize a care plan. This tool is mandated for use in LTC facilities by the Health Care Finance Administration. The primary purpose of the MDS is to provide information to decision makers that will lead to ways to improve the care of residents in LTC facilities through comprehensive assessment and informed care planning.

Setting and Sample: The targeted population is the elderly residents of LTC facilities during the year 1996. The data set was constructed by including all residents aged between 65 and 100 years old. In addition, they must have had a first initial admission assessment, an initial comprehensive assessment, or a readmission assessment. Finally, they must also have had at least 2 records and have a documented "history of falling within 30 days" (which is also the outcome variable in the later record). The total number of resultant cases in this data set was 9980.

Procedure.

1. After obtaining approval from the Institutional Review Board (IRB) at the University of Wisconsin - Madison, the MDS data set was obtained from the Center for Health Systems Research and Analysis (CHSRA), a research center at the University of Wisconsin-Madison. To protect the privacy of individuals, CHSRA removed all personally identifying information, including the person's name, Social Security Number, Medicare number, Medicaid number (if any), and birth date.
2. To answer whether the fall models constructed using the LBP data mining technique can select the important variables correctly, we must supply the model with both variables already known to be associated with falls and other variables that are not. All of the models constructed using the LBP technique were supplied with reduced data sets of five or six variables. Some variables were known to be associated with falls. Others were known NOT to be associated with falls. Each reduced set was

constructed to contain both types of variables. In this case, the known variables were those specified in the Resident Assessment Protocols (RAPs),[9] which is the process used to assess nursing home residents who have problems (such as falls or incontinence) that have an especially significant impact on their care. For this study, the model was given four variables known to be associated with falls. These variables were; *fell in last 30 days*, *fell in last 31 – 180 days*, *antipsychotic* (received antipsychotic medication in last 7 days), and *hemiplegia or hemiparesis*. The model was also given two variables known not to be associated with falls; *hearing problem* and *mode of expression: writing*.

3. The LBP technique was employed to determine whether a model could correctly select the variables associated with falls. If the LBP method is working properly, then it should identify as "important" only those variables known to be associated with falls. To determine which of the supplied variables were important, the LBP models calculated the L_1 norm of error when each variable was removed. The variables with an L_1 norm (the sum of absolute errors derived from the LASSO shrinkage procedure of the Basis Pursuit Method.)[5] greater than the threshold value (0.1) were considered to be important. The associated variables selected by the models were then matched against the variables contained in the RAPs. Once we were confident the LBP technique could properly identify the variables associated with falls, a model that provided the likelihood of falls among the elderly in long–term care facilities was generated.

.

RESULTS

The LBP techniques used to construct this model correctly identified as important three of the variables known to be associated with falls — *fell in last 30 days* (L_1 norm = 0.918), *antipsychotic* (L_1 norm = 0.601), and *hemiplegia or hemiparesis* (L_1 norm=0.127). It also identified the two unimportant variables correctly: *mode of expression: writing* (L_1 norm = 0.015) and *hearing* (L_1 norm = 0.007). However, the LBP technique incorrectly identified *fell in last 31 – 180 days* (L_1 norm = 0.054) as an unimportant variable.

We carried out further analysis to determine the effect of the relationship between predictive variables on the performance of models constructed using the LBP data mining technique. Two variables, *fell in last 30 days* and *fell in last 31 – 180 days*, are highly correlated (p-value < 0.01). If the *fell in last 31-180 days* variable was removed from the model, the three remaining variables known to be associated with falls were identified as above. But, the L_1 norm of each variable changed slightly as follows: *fell in last 30 days* (L_1 norm = 0.915), *antipsychotic* (L_1 norm = 0.605), and *hemiplegia or hemiparesis* (L_1 norm = 0.113). It also identified the two unimportant variables correctly: *mode of expression: writing* (L_1 norm = 0.000) and *hearing* (L_1 norm = 0.007).

Furthermore, if the *fell in last 30 days* variable was removed and replaced with the *fell in last 31-180 days* variable, the model also correctly identified the new set of three variables known to be associated with falls. The three important variables were identified as *fell in last 31-180 days* (L_1 norm = 0.100), *antipsychotic* (L_1 norm = 0.606), and *hemiplegia or hemiparesis* (L_1 norm = 0.190). It also identified the two unimportant variables correctly: *mode of expression: writing* (L_1 norm = 0.000) and *hearing* (L_1 norm = 0.036)

Probabilities were calculated using the model with three variables known to be associated with falls: *fell in last 30 days*, *antipsychotic*, and *hemiplegia or hemiparesis*. The exact probability of falling within the next three months was calculated for each individual based upon the specific combination of variables unique to that individual. For example, the model calculated that the residents who had the values of all three of these important variables equal to 0 had a baseline probability of fall of 0.27. This means that the elderly residing in LTC facilities who did not have a history of falls, did not received any antipsychotic drugs, and did not have either hemiplegia or hemeparesis would have a likelihood of falling of 0.27.

The results further showed that residents who had only *antipsychotic* had a probability of fall ranging from 0.321 to 0.482. The probability of falls is related to the number of days that the residents received antipsychotic drugs. However, the relationship is not linear as the residents who received only antipsychotic drugs had a probability of fall ranging from 0.321 to 0.482. The probability depended upon how many days the resident received the drugs within the last 7 days. Up through the fourth day of receiving the drugs, the probability of fall increased each day. After the fourth day, the probability trended back downward each day.

While the presence of *hemiplegia or hemiparesis* does have a significant effect on the probability (it has a larger L_1 norm), this effect is "negative" in that the presence reduces the likelihood of falling within the next three months. Conversely, this indicates that the absence of *hemiplegia or hemiparesis* actually increases the probability of falls among residents. For example, if the resident received antipsychotic drugs for at least one day during the period, *fell in last 30 days*, but did not have *hemiplegia or hemiparesis*, the probability of fall was 0.76. However, if the resident had all three variables, including *hemiplegia or hemiparesis*, the probability of fall decreased to 0.71.

According to this model, the resident with the least likelihood of falling is one who has only *hemiplegia or hemiparesis* (0.219). The greatest probability of fall

occurs amongst those who received antipsychotic drugs for 4 days out of the last 7, *fell in last 30 days* and had no *hemiplegia or hemiparesis* (0.854).

Previous studies have indicated that, when using LBP, the addition of "unimportant" variables into the model building process has little effect on the probabilities. Because *fell in last 30 days* is highly correlated with *fell in last 31-180 days*, the LBP technique flags one of them as unimportant. However, given the results of other studies, we would not expect this to change the overall results much. Thus, we would expect that the probability of fall for those who had only *fell in last 30 days* is very similar to those who have both *fell in last 30 days* and *fell in last 31-180 days*. To test this, we constructed a second model using both variables. This second model confirmed that the results were very similar. For example, residents who had only *fell in last 30 days* had a probability of fall of 0.70. If the resident had *fell in last 30 days* and had also *fell in last 31-180 days*, the probability slightly increased to 0.72.

DISCUSSION

In this study, we ran preliminary tests using the KDD process along with a new data mining technique, Likelihood Basis Pursuit (LBP), to construct models that could predict the likelihood of falls. Because the model output provides probabilities based upon the exact combination of variables present in a particular resident, models constructed using the LBP technique have the potential to be more useful than classification models for assessing fall risk. While this study only tested the LBP technique with six variables, some interesting results were found nonetheless. First, we found that when given a mix of variables—some of which were associated with falls and some not—the LBP technique can select the variables that the literature identifies as associated with falls [9,10] correctly, provided no association exists among the predictive variables. Furthermore, the models constructed using the LBP technique have the capacity to identify the probability of falling within the next three months given a particular combination of these selected variables. Thus, they are able to tailor their predictions in a highly individual way. The probability results were consistent with our knowledge that falling is caused by multiple variables. Each variable had different effects on the likelihood of falling. Some variables, such as *fell in last 30 days* and *antipsychotic*, increased the probability of falling. Others, such as *hemiplegia or hemiparesis*, decreased the risk of fall. Even though there is no previous literature supporting this finding, one explanation might be that the elderly who have hemiplegia might receive closer care and might not move without assistance from others. Thus, a resident who has *hemiplegia* combined with *fell in last 30 days* and *antipsychotic* had a greater probability of falling than a resident who had only *hemiplegia*, but a lower probability than a resident who had only *fell in last 30 days* and *antipsychotic*. The ultimate probability that an individual would fall depended upon the specific combinations of these variables.

We also observed that, even though the probability of falls is related to the number of days that the residents received antipsychotic drugs, the relationship is not linear. Residents who received only antipsychotic drugs had a probability of fall ranging from 0.321 to 0.482, depending upon how many days the resident received the drugs within the last 7 days. The probability of fall increased up through the fourth day of receiving the drugs, after which the probability trended back downward. After closer examination of the data, we found that only a few residents received antipsychotic drugs for fewer than 6 days. The increased fall rate among residents who received antipsychotic drugs until the fourth day might reflect a physiological adjustment period to the action of medication and their underlying disease.

This is a real finding, not an anomaly due to a low prevalence rate. The LBP probability calculations are probably accurate in regimes where the frequency rate is very low. The LBP probability calculations can even fill gaps in regimes where the frequency rate is zero. [5,6]

This study also demonstrated that associated variables produced nearly identical probabilities of fall. For instance, *fell in last 30 days* and *fell in last 31-180 days* are very strongly correlated with each other. Thus, results were very similar for those who had only *fell in last 30 days* compared to those who had both *fell in last 30 days* and *fell in last 31-180 days*. This indicates that if we have variables that are correlated with each other, we can reduce the number of variables used to construct the model by selecting the strongest variable from the set of correlates.

If they were considered on their own, *fell in last 30 days* and *fell in last 31-180 days* were both considered to be important variables by the model. However, if they were combined together, the model would identify *fell in last 31-180 days* as unimportant. This supports professor Wahba's comments that the correlations among variables would affect the performance of the LBP data mining technique. In this study, if two highly correlated variables known to be associated with falls were included in the model, the LBP technique would identify only one of the two as important to falls. Thus, under these circumstances, the model might exclude a variable that was, in fact, very important on its own.

This study shows that the KDD process using the LBP data mining technique can provide a more useful model output. The predictive power of the probabilities supplied by the LBP technique may be of greater

clinical value than classification. The LBP technique provides output in terms of a probability that an individual will fall based upon the specific combination of variables unique to that individual. On the other hand, other data mining techniques such as Support Vector Machine (SVM) provide output in terms of classification (whether the elderly would fall or not fall) based upon a specific weighted combination of variables. This combination is fixed. This might not be useful in practice because it requires that health care personals know the status of all of the variables in a particular resident in order to determine a classification. But, the status of all of the variables might not be available. If information about a single important variable is missing, then the classification will not be reliable. Because LBP provides a probability based on flexible combinations of variables, residents for whom the status of only a few variables are known can still be assessed as to their probability of falling. Furthermore, because the effect of each variable is much more clear, each variable can be treated in isolation. Because the information is much more specific, a much more tailored intervention strategy could be applied to those at risk.

While a model that provides explicit probabilities of fall based upon combinations of variables does not give such a simple yes/no answer, it does provide an opportunity for health care providers to integrate their judgment, based upon experience, into fall prevention. If multiple risk factors are presented, health care providers must still decide which risk factors to address first when providing fall prevention intervention. For instance, if we know that elderly residents receiving antipsychotic drugs have a greater likelihood of falls than those who do not, health personnel will know to pay closer attention to those receiving the medication and to arrange for more oversight from caregivers. Furthermore, the decision to provide antipsychotic drugs should be made in full consideration of this risk.

As it stands, this method does have limitations. One area that needs further investigation is the fact that models constructed using the LBP technique require little correlation exist among the predictor variables. While prediction accuracy is probably not affected by such correlations, the identification of the variables having the most influence on the predictions can be dramatically changed by correlations. It is a given in health related data that many of the variables are correlated with each other. Other techniques, such as Principal Component Analysis (PCA) or Factor Analysis, might need to be employed to eliminate the correlation in order for these models to succeed.

Further investigation is also needed into the question of how many variables LBP can handle simultaneously before the LBP technique can be employed with confidence. This study used only five or six variables with the LBP technique. Most nursing phenomena are caused by multiple variables and it is clear that models will require the ability to handle many variables if they are to successfully model health phenomena. The LBP technique remains untested on larger health related data sets. Thus, the LBP technique will also require further study before it can be applied in a practical health care environment.

Acknowledgement

The authors are also grateful to Dr. Grace G. Wahba, Dr. Hao Zhang, Dr. Meta M. Voelker,, and Ya-Ju Fan for their help with implementing the LPB algorithms and the associated programming tasks. Furthermore, I would like to acknowledge the Center for Health Systems Research and Analysis for its assistance on the MDS database.

Reference

1. Berger AM, Berger CR. Data mining as a tool for research and knowledge development in nursing. Comput Inform Nurs. 2004 May-Jun; 22(3):123-31.
2. Lee IN, Liao SC, Embrechts M. Data mining techniques applied to medical information. Med Inform Internet Med. 2000 Apr-Jun; 25(2):81-102.
3. Goodwin LK, Iannacchione MA, Hammond WE, et al. Data mining methods find demographic predictors of preterm birth. Nurs Res. 2001 Nov-Dec;50(6):340-5.
4. Abbott PA, Quirolgico S, Candidate D, et al. Can the US minimum data set be used for predicting admissions to acute care facilities? Medinfo. 1998;9 Pt 2:1318-21.
5. Zhang H, Wahba G, Lin, et al. Variable Selection and Modeling Building via Likelihood Basis Pursuit. Technical Report no. 1059. July 9, 2002. Madison: University of Wisconsin-Madison.
6. Wahba G, Wang Y, Gu C, et al. Smoothing spline ANOVA for exponential families, with application to Wisconsin Epidemiological Study of Diabetic Retinopathy. The Annals of Statistics. 1995 1865 – 1895.
7. Rubenstein LZ, Josephson KR, Osterweil D. Falls and fall prevention in the nursing home. Clin Geriatr Med. 1996 Nov;12(4):881-902.
8. Thapa PB, Brockman KG, Gideon P, Fought RL, Ray WA. Injurious falls in nonambulatory nursing home residents: a comparative study of circumstances, incidence, and risk factors. J Am Geriatr Soc. 1996 Mar;44(3):273-8.
9. Boulter CS, Morris JN, Hawes C, et al. Minimum Data Set Reference Manual. Natick: Eliot Press, 1993.
10. Tideiksaar R. Falling in Old Age. New York: Springer Publishing Company, 1996.