34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Measuring MERCI: Exploring Data Mining Techniques for Examining the Neurologic Outcomes of Stroke Patients Undergoing Endo-vascular Therapy at Erlanger Southeast Stroke Center*

Matthew McNabb, Yu Cao, Ph.D., *Member, IEEE*, Thomas Devlin, M.D., Ph.D., Blaise Baxter, M.D., Albert Thornton

***Abstract*— Mechanical Embolus Removal in Cerebral Ischemia (MERCI) has been supported by medical trials as an improved method of treating ischemic stroke past the safe window of time for administering clot-busting drugs, and was released for medical use in 2004. The importance of analyzing real-world data collected from MERCI clinical trials is key to providing insights on the effectiveness of MERCI. Most of the existing data analysis on MERCI results has thus far employed conventional statistical analysis techniques. To the best of our knowledge, advanced data analytics and data mining techniques have not yet been systematically applied. To address the issue in this thesis, we conduct a comprehensive study on employing state of the art machine learning algorithms to generate prediction criteria for the outcome of MERCI patients. Specifically, we investigate the issue of how to choose the most significant attributes of a data set with limited instance examples. We propose a few search algorithms to identify the significant attributes, followed by a thorough performance analysis for each algorithm. Finally, we apply our proposed approach to the real-world, de-identified patient data provided by Erlanger Southeast Regional Stroke Center, Chattanooga, TN. Our experimental results have demonstrated that our proposed approach performs well.**

I. Introduction

MERCI [1] is a relatively new medical procedure released by the Food and Drug Administration in 2004, which widens the therapeutic window for removing deadly blood clots from the brain to 8 hours after the onset of acute ischemic stroke (AIS). The only additional treatment for acute stroke, Tissue Plasminogen Activator (tPA), is a drug approved in 1996administered intravenously (IV) for dissolving clots within the brain, benefits approximately 1 in 8 patients and has FDA approval for only 3 hours after stroke onset. Although associated with improved morbidity (i.e., patients functionality and quality of life), IV-tPA did not improve mortality in pivotal trials [2]. In addition, only a small fraction of patients who qualify for IV-tPA actually receive this treatment [2]. This expansion of the time window offered by MERCI can be critical to the patient's outcome and may be particularly beneficial in patients with large clot burden [3]. While the usage of MERCI, and two other clot extractors currently approved for thrombectomy in patients with AIS [4, 5], is rapidly growing, the lack of placebo controlled trials in the stroke device field and the variability of study results has contributed to hesitation on the part of treating physicians to fully embrace this new technology. Furthermore, the complexity of variables relating to the medical treatment of these patients makes study result interpretation and comparison challenging. As part of a large national registry [5], Erlanger Hospital of Chattanooga, Tennessee and the University of Tennessee: College of Medicine Chattanooga (UTCOMC) has collected data strictly confined to stroke patients treated by MERCI, including a generous collection of detail with regards to procedure analysis, including patient diagnosis and status levels coming in the door, surgical procedure data, patient status after the procedure, and 90-day follow-ups [6]. In this paper, by introducing data mining techniques to this study, we ultimately hope to not only reinforce the findings of conventional statistical approaches, but to search for new relationships and significance within the data that might not have been found before.

To this end, we are looking to successfully mine the data collected by Erlanger hospital to produce effective prediction weights, which should reasonably predict a new patient's outcome. Using transparent methods for this prediction demonstrates easily recognizable relationships with the data.

Much larger collections of data exist in state or national stroke registries, but such sources are often difficult to access for purposes in individual center academic research. Our smaller collection of data provided by Erlanger represents a specific study of a more homogenous patient population from the south. Furthermore, working with a smaller data set with many attributes per instance is an issue that we have not seen addressed as frequently in our search for other data mining studies, and a new study on this subject might open another niche of discourse in Information Science on solving some real problems [7, 8].

*Resreach was supported in part by the National Science Foundation (NSF) under Grant IIS-1156639, Grant DBI-0821820, and by the State of Tennessee.

Matthew McNabb and Yu Cao are with College of Engineering and Computer Science, The University of Tennessee at Chattanooga, Chattanooga, TN, 37403 (corresponding author to provide phone:423-425-4351; Fax:423-425-5442; e-mail: yu-cao@utc.edu).

Thomas Devlin, M.D, Ph.D., and Blaise Baxter, MD, are with the Erlanger Southeast Regional Stroke Center, the University of Tennessee: College of Medicine Chattanooga (UTCOMC), Chattanooga, TN, 37403 USA (e-mail: tgdevlin@bellsouth.net).

Albert Thornton is with Troy University, Alabama. He has contributed to this paper as part of his NSF REU summer research internship at UTC.

II. DATA

A. Overview of Data

The specific study used for this data mining survey consists of a total of 115 patients meeting the following criteria: Each patient was diagnosed with acute ischemic stroke; Arrival for treatment was between 3 and 8 hours after stroke onset; Each patient was over 18 years of age; Each patient exhibited an ischemic stroke clinically with CT neuroimaging disclosing hypodensity less than 1/3 the middle cerebral artery territory. Because this data set is exclusively made up of MERCI patients, our study is specifically limited to gauging factors for patient health and mortality given MERCI treatment. Therefore, without a strict control group our goal becomes not that of measuring the effectiveness of MERCI itself in relation to other treatments or little to no treatment at all, but rather a search for relations among more specific factors within the MERCI process itself. This level of study is an appropriate stage of specificity progression, given that medical trials have already established MERCI to be "cost effective" compared to other procedures within its approved treatment window. An evaluation of more specific details within a more developed branch of the decision making process further sharpens our understanding of information closer to the end result. Our goal is specifically to create a successful automated prediction system, which predicts the outcome of subsequent input data of a similar sort, and is easily discernible by analysis both inside and out of the Information Science discipline to determine the most significant prediction sources.

B. Choose Attribute

The detail size of this sample set provides for a variety of choices from which to draw statistical trends. These details can be categorized a number of ways, but for machine learning and data mining algorithms, the following categories split the data according to their logical functions in discovery: (1) Static data – Information that does not involve any choice in the medical process. For example, Personal information: age, gender, etc, Diagnosis information, Location of clot; (2) Non-static data – information that could be effected by decisions in treatment. For example, Procedural information, Device usage, Onset to puncture, Procedure length; (3) "Negligible" or repetitive institutional data – Some details are simply listings of facilities, patient numbers, surgeons, etc. It could be argued that not all of these are irrelevant, but we have discarded them for this study, either by negligibility, or due to repetition in values; (4) Outcomes and post-procedure data. Ultimately, 40 useful instance attributes were identified for input. Of the attributes identified as possible outcome gauges, 2 were chosen for class attributes: 90-day mortality, a check for patient survivability 90 days after the endovascular procedure with MERCI, and 90-day MRS (Modified Rankin Scale), a more precise measurement of the patients neurologic recovery.

C. Sample set size

Because 3 patients had no result attributes recorded, they were eliminated from the data set, leaving 112 instances for training and testing. Many modern data mining efforts use massive amounts of data to sharpen weights and render detail size maxima negligible for their algorithms, so this number could be said to be small by relation, although it represents a great deal of gathering work and is over three times the size of the typical medical trials examined in preliminary research. While hundreds of general stroke examples might be obtained nation-wide via a stroke registry, this more specialized set emphasizes the more exact details of a much smaller sub-group in a local area. Thus, new samples are more difficult to come by, but the set provides more specialized trends that would likely not be conclusively found in more general databases without sifting with correlative search criteria. In any event, this set is still large enough to establish trends, but the marginal return of all 40 attributes may be limited due to insufficient instances to train them.

D. Balance

In terms of 90-day mortality, the data was well balanced, with 53 deaths and 59 survivors. The 90-day MRS benchmark, however, is unbalanced by nature. Figure 1 shows the distributions of ratings 0 – 6. As a set of numbers, this gauge is heavily unbalanced to a rating of 6, and drives most classifiers to overestimation. When an understanding of its semantics is applied, however, the skewed weight towards 6 is simply a result of 6 representing death. Life, then, is divided among ratings 0-5. If possible, it would be beneficial for information science to be able to divide death into similar sub-categories, to represent varying levels of how close each patient's demise was to recoverability. Perhaps one patient was completely un-savable, regardless of any intervention at all, but another might have been saved with different choices made. This falls into the realm of speculation, however, and death is, in reality, a condition without levels – dead is dead, and one cannot become better or worse after dying. Nonetheless, applying the fact that MRS = 6 refers to death, this apparently unbalanced data set rather becomes a two-layered classification problem with sub-problems of a much more balanced nature. Without rating 6, while we have no instances of 5 or 0, the distribution still has a somewhat reasonable bell-curve shape. If the data set is first analyzed in terms of life and death, those instances that received a life prediction could be further analyzed with weights trained to predict MRS ratings 0-5. The second layer, then, would be trained with the 59 survivors

III. PROPOSED APPROACH

A. Choosing Classifiers

Ultimately, a variety of classifiers should be used for an exhaustive study of this data, since automated test processes can run day and night for further analysis. For this stage in examining the data, however, we were most interested in a more limited set of criteria. *"White box" classification*: Ideally, we would like the end results to be easily

understandable by analysts outside of the Information Science discipline. Multi-layered approaches, such as neural networks, use a system of derived weights which may train well, but do not present a clear and easily traceable line of significance back to the original attributes. Single-layered weights can convey the significance of each attribute more plainly, and are thus encouraged. This exclusion of machine learning complexity, of course, may eliminate algorithms that might perform better, but similar complexities can be added with more easily traceable layering, as will be explained later in Attribute Layering. *Initial Naïve Analysis*: Initially, we would like our algorithm to make no automatic assertions that any attributes are related to one another, allowing weights to be derived as independently as possible. In the end, however, this will likely not be the optimal solution. We know already that major groupings of attributes exist: patient vitals, procedure data, and post-procedure data. It is our initial assumption that none of the inputs for these experiments are reliant upon the other. This, of course, may not ultimately be appropriate for some sub-sets of inputs, such as groupings of instruments used, but any relationships to these can be inserted afterwards with Attribute Layering. *Flexibility*: Technology that handles missing data is a must, as each instance sample is highly valued. It is also preferable, while single-layered algorithms are preferred for their transparency, that these algorithms can be easily stacked to address our balancing issue with the 90-day MRS result. A number of data mining methods were considered in preliminary experiments at the beginning of the study, with emphasis on linear regression and Bayesian techniques which satisfied the above characteristics. At the beginning of our study, we began manual experimentation with several data mining approaches, using the University of Waikato's Weka GUI, to find a suitable candidate for programmatic refinement. Logistic Regression proved to be the best performing approach meeting the above criteria in the experimental stage. For this reason, coding for the remainder of our proposed approach was tested and debugged using it as a prime algorithm. More success in the manual experiment stage does not, however, indicate that Logistic Regression will be the best choice once all other parts of our solution are implemented, so a return to other algorithms is far from out of question in future work.

B. Searching for Significant Attributes

Although Logistic Regression proved to be the best of the white-box methods evaluated by initial experimentation, it suffers from a tendency to overestimate weights in data sets where the number of detail variables is relatively large with respect to instance samples provided [9]. This tendency is in the nature of variable estimation, particularly demonstrated in Gaussian Elimination solutions to algebraic systems by "free variables."

Unlike the equations of a linear system in a mathematics textbook, a data set is an abridged representative of its universe – if we do not have enough samples to pin down the significance of each attribute provided, the significance of every attribute can be skewed on the whole [9]. Therefore, we need to develop a new approach to choose a combination of attributes which best affects prediction performance. Four methods of doing so were examined, with varying reliance on estimated weights.

Depth First Search (DFS): A DFS approach was employed to exhaustively search every possible attribute combination for the most effective. With 40 input attributes to choose from, the complexity of this algorithm becomes a choosing equation. While this may ultimately find the best combination given enough time, the complexity of the search is explosive as more attributes are added [10]. Because an immense runtime would be required for a complete DFS for all attribute combinations, a pre-specified attribute limit parameter was put in place to keep the algorithm's runtime within several hours. Some technological approaches, such as multiprocessing or more robust and/or dedicated machines could be used to make DFS faster, but this incredible complexity in any case requires substantial computations to finish;

Naïve Smart Search: To improve on DFS, our simplest approach uses a saved-progress function related to dynamic programming. As with DFS, the number of attributes to be chosen is provided beforehand, in order to divide the problem into batch chunks and simplify the logic. Given a pre-chosen attribute selection r, the algorithm begins with a simple selection of the first r attributes in the set, calculating their score. For example, representing attributes with numeric names for simplicity, with a given $r=5$ and an initial prediction success rate of 43%. Being the first evaluation made, this score is saved as the "best" evaluation. The search continues by examining the 6^{th} attribute. An evaluation is made of r sets of attributes, with the new attribute replacing one of the attributes in the current "best" evaluation. After the accuracy of each new evaluation is calculated, all $r+1$ (in the case of our example, 6) evaluations are compared, and the best of those becomes the new "best." This process is repeated until all attributes have been processed. Given 1 combination for the first r attributes, followed by r combinations for each of the remaining $n-r$ attributes, we have $r(n-r)+1$ complexity for each r, resulting in the following complexity for calculating all r. By eliminating attributes that perform less accurately, Smart Search runs with a complexity of $2.81*10^{-7}$ percent of that required by DFS. Apart from using no direct heuristic to further eliminate search branches, the disadvantage of using this algorithm to build a combination set, starting from attribute 1 and traversing to attribute n, is that, by design, many correlations between attributes from one side of the set to the other are eliminated from evaluation;

Weight-based Search: The particularly crippling reality for any attribute choosing algorithm is that each unit of complexity represents a training session, which is expensive in its own right. With 112 examples, our data set's training session will require a relatively short period of machine time, but as the dataset grows, further elimination of complexity may play a key role in investigating larger data sets, which may not have reached the point yet where attribute number is negligible and attribute finding algorithms are depreciated. Naturally, the magnitude of weight assigned to each attribute is a measure of that attribute's estimated significance to the outcome of the current combination. Thus, it is a natural and

automatic heuristic for eliminating least effective attributes – this eliminates the need to evaluate each attribute in a combination as a drop candidate, reducing complexity by a factor of *r*, if following the pattern of Naïve Smart Search. While this application significantly reduces the complexity of Naïve Smart Search in its steps, it still retains *O(n^2)* complexity in its upper bound. Carrying reliance on the heuristic further, it would be possible to begin with all 42 attributes, eliminating them one at a time according to the lowest weight, until the outcome of the machine's prediction ceased to improve, coming to an estimated set of "most significant attributes" in *O(n)* time.

Weight-Based Smart Search: To take full advantage of a weight-based heuristic with a *1* to *n* algorithm, the summation property of Naïve Smart Search can be eliminated by adding and removing attributes on the fly. The state machine in Figure 2 illustrates three different phases for adding and removing attributes in a Weight-Based Smart Search. The algorithm begins by adding attributes and evaluating Logistic Regression after each add. As long as the performance of our combination improves, we continue adding.

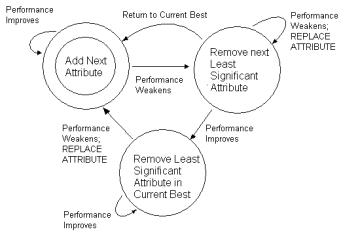

Figure 2: State machine that illustrates three different phases for adding and removing attributes

IV. Results and Conclusions

Due to constraints of time, a complete measurement of attribute finding with Depth First Search was impossible. Searching with a constraint of 3 attributes for single-layered Logistic Regression required 8 hours on an i7 system with 1 GB dedicated to the heap, yielding a success rate of 63% for 90-day mortality training. A constraint of 4 attributes required over 24 hours and did not complete. Naïve Smart Search was able to reduce runtime for a single attribute constraint to minutes, and completed the entire batch of constraints in 40 minutes, with a success rate averaging at roughly 74% in single-layer Logistic Regression. This performance was consistent, with waivering of some +-1.5% success rate due to the random number generators used in the training algorithms. False Negatives vs. False positives were balanced, with the following attributes selected as significant: Age, Intubulation, IV Lytic, Length of Procedure, Diastolic B/P, Right Anterior (location of clot), ICA, MCA, IA Lytic, Count of V2.5 Firm. Top-down weight-based attribute choosing, while finishing very quickly in a few seconds, only managed to yield a 63% success rate. As this algorithm is not novel to this project, it might be said to be a control on the other end of the spectrum to DFS. Weight-based Smart Search, executing in less than 30 seconds, yields a success rate averaging around 70%. Figure 4 shows a graphic representation of all four algorithms. The success rate of DFS is a guess, since a full iteration of it could not be completed. Initial MRS performance, using conventional prediction without splitting the problem, was at 42%, with similar attributes selection dwelling more on the location of the clot. As was the case in initial trial experiments, conventional MRS prediction leans heavily towards rating 6. The patient's diagnosis and vital signs played the largest role in our prediction criteria, but procedural data also played a part. It is likely that a hypothetical dataset containing a control group of unfortunate patients, for whom no treatment was available, would see more reliance on procedural data. Making a distinction between the impact of differences in medical choices and the differences in patient condition which influenced those choices is still up to Medical Science at this point. While the goal included advancements towards medical predictions, the fruits of this research are more generally applicable to Information Science as a whole. There is little about these solutions that are particular to medical data mining – machine learning for any dataset with the same concerns for data Instances vs. data attributes can benefit from attribute choosing algorithms.

The authors would like thank the insightful comments from Peter Catalano during paper writing. Peter directs administration and development for the Pleiades Foundation for Advanced Neuromedical Education, Chattanooga, TN.

V. References

[1] A. A. Khalessi, S. K. Natarajan, D. Orion, M. J. Binning, A. Siddiqui, E. I. Levy, and L. N. Hopkins, "Acute Stroke Intervention," *Journal of the American College of Cardiology: Cardiovascular Interventions,* vol. 4, pp. 261-269, 2011.

[2] "Tissue Plasminogen Activator for Acute Ischemic Stroke by The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group," *The New England Journal of Medicine,* pp. 1581-1588, 1995.

[3] E. Barker, "A new weapon to combat stroke," *Modern Medicine,* vol. 69, pp. 26-29, 2006.

[4] "The penumbra pivotal stroke trial: safety and effectiveness of a new generation of mechanical devices for clot removal in intracranial large vessel occlusive disease," *Stroke,* vol. 40, pp. 2761-2768, 2009, written by Penumbra Pivotal Stroke Trial Investigators.

[5] R. J. JL Saver, E Levy, T G Jovin, B Baxter, R Nogueira, W Clark, R Budzik, OO Zaidat, "Primary Results of the SOLITAIRE™ FR With the Intention for Thrombectomy (SWIFT) Multicenter, Randomized Clinical Trial," in *Proc. of International Stroke Conference*, New Orleans, Louisiana, U.S.A, 2012.

[6] T. Devlin, B. Baxter, T. Feintuch, and N. Desbiens, "The Merci Retrieval System for Acute Stroke: The Southeast Region Stroke Center Experience," *Neurocritical Care,* vol. 6, pp. 11-21, 2007.

[7] N. Thai-Nghe, Z. Gantner, and L. Schmidt-Thieme, "A New Evaluation Measure for Learning from Imbalanced Data," in *The 2011 International Joint Conference on Neural Networks (IJCNN)*, San Jose, CA, 2011.

[8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques, Third Edition*. San Francisco, CA, USA: Morgan Kaufmann, 2011.

[9] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*. San Francisco, CA, USA: Morgan Kaufmann, 2011.

[10] J. H. Klotz. (2006). *A Computational Approach to Statistics*. Available: http://www.stat.wisc.edu/~klotz/Book.pdf