Mortality Prediction using Similarity Measures for Medical Event Sequences

Completed Research

Joel Fredrickson  
University of Colorado Denver  
Joel.Fredrickson@ucdenver.edu

Michael Mannino  
University of Colorado Denver  
Michael.Mannino@ucdenver.edu

Omar Alqahtani  
University of Colorado Denver  
Omar.Alqahtani@ucdenver.edu

Farnoush Banaei-Kashani  
University of Colorado Denver  
Farnoush.Banaei-Kashani@ucdenver.edu

Abstract

We extend a similarity measure for medical event sequences (MESs) and evaluate its performance on mortality prediction using a substantial trauma data set. We extend the Optimal Temporal Common Subsequence for MESs (OTCS-MES) measure by generalizing the event-matching component with user-defined weights. In the empirical evaluation of classification performance, we provide a more complete evaluation than previous studies. We compare the predictive performance of the Trauma Mortality Prediction Model (TMPM), an accepted regression approach for mortality prediction in trauma data, to nearest neighbor algorithms using similarity measures for MESs. Using a data set from the National Trauma Data Bank, our results indicate improved predictive performance for an ensemble of nearest neighbor classifiers over TMPM. Our analysis demonstrates a superior Receiver Operating Characteristics (ROC) curve, larger AUC, and improved operating points on a ROC curve. Predictive performance improves for the ensemble for a variety of sensitivity weights and false positive constraints.

Keywords

Medical event sequences, trauma mortality prediction, similarity measure

1. Introduction

This study involves mortality prediction for trauma centers, an important classification task with established methods. Trauma injuries account for a substantial number of deaths as the leading cause of death in people younger than 44. In addition, trauma is the fifth leading cause of death for all age groups (Glance et al. 2009). Treatment methods for trauma related injuries are extremely costly, often leading to expensive forms of care. Cassidy et al. (2014) explains “accurate injury severity scoring systems are essential for benchmarking outcomes and objectively evaluating and improving trauma care.”

Evaluating trauma care based on guidelines for severity dependent mortality rates involves retrospective mortality prediction. Typically, mortality prediction models use information from historical trauma incidents to correlate patient attributes and injury severity to known trauma discharge dispositions (deceased or non-deceased). These retrospective mortality prediction methods have “important clinical and economic implications because these tools are used to evaluate patient outcomes and quality of care” (Weeks et al. 2016).

Because of the importance of mortality prediction for trauma centers, researchers have developed several prominent prediction methods. The most widely accepted method, the Trauma Mortality Prediction Model (TMPM), involves detailed regression modeling of individual injury codes using a large training sample. TMPM (Glance et al. 2009) uses derived coefficients for more than one thousand injury codes to make mortality predictions.
In this paper, we study an alternative approach to mortality prediction based on similarity of medical events in a patient’s trauma incident. Predicting mortality based on similarity provides better explanation than regression prediction as similar cases provide explanation of a prediction. Prediction based on similarity using nearest neighbors classification does not require training although it requires indexing of trauma incidents for efficient computation of nearest neighbors.

A similarity measure for medical event sequences (MESs) is important for many reasoning tasks and practical applications useful to health care professionals and data mining algorithms. Medical data warehouses contain large volumes of MESs, state sequences relevant to health care. Health care professionals can use a similarity measure to develop treatment plans based on similar patients. Data mining algorithms for risk assessment, condition identification, and conformance to clinical pathways can use a similarity measure. Furthermore, research opportunities about similarity measures exist for patient classification into morbidity or risk groups, evaluation of patient adherence to a specific care management plan, and discovery of similar patients for medical social networking.

In a prior research study (Mannino et al. 2017), we developed the Optimal Temporal Common Subsequence for Medical Event Sequences (OTCS-MES). In this study, we generalize event matching in the OTCS-MES with user-defined weights. For mortality prediction in trauma data, we use an event severity weight in addition to event prevalence. Generalization of weights for event matching is important for using OTCS-MES similarity measures in a wider variety of medical domains.

We compare predictive performance of nearest neighbor classification using MES similarity measures to TMPM. In nearest neighbor classification, we use the OTCS-MES measure with two weighting approaches (event prevalence and event severity), the original OTCS with only exact matching of event codes, and an ensemble using these three classifiers. We compare performance for three important performance measures (receiver operating characteristic (ROC) curves, area under a ROC curve (AUC), and operating points derived from a ROC curve) using a data set from the National Trauma Data Bank. Our results indicate superior performance for an ensemble of nearest neighbor classifiers over TMPM on ROC curve analysis and AUC. For optimal operating points, the ensemble provides better performance than TMPM especially as the importance of sensitivity increases.

This study makes three important contributions. Most importantly, this study provides a new classification method with better performance than the accepted standard, TMPM. The ensemble of nearest neighbor classifiers obtained better performance than TMPM on ROC curves, AUC, and optimal operating points on a ROC curve. As an important secondary contribution, generalization of the event-matching component of the OTCS-MES measure makes event matching applicable to a wider variety of medical domains and decision-making tasks. As another secondary contribution, the detailed performance comparison provides a more complete analysis than previous studies. Prior studies on mortality prediction neglected to compare operating points on a ROC curve. No studies have used nearest neighbor classification for mortality prediction with an uncommon mortality class.

This study continues as follows. The second section presents the design of the experiment comparing nearest neighbor prediction using the revised OTCS-MES and TMPM. The third section presents results of the experiment and discusses implications. The last section summarizes the paper and identifies future extensions.

2. Design of Empirical Evaluation

This section provides details about the design of an experiment to compare nearest neighbor algorithms using similarity measures to TMPM on mortality prediction. The experiment design covers presentation of treatments involving similarity measures and nearest neighbor algorithms, research questions and hypotheses, data, and performance measures.

2.1 Similarity Measures

This study uses three similarity measures, the original OTCS, the OTCS-MES with event prevalence weights (OTCS-MES EP), and the OTCS-MES with event severity weights (OTCS-MES ES). Although all three measures contain components for event matching and temporal structure of events, this study only uses the event matching component because trauma records do not have a temporal structure.
2.1.1 Original OTCS

The original Optimal Temporal Common Subsequence (OTCS), developed by Zheng et al. (2010), uses exact matching for events. Given a state-sequence defined as \( S_n = [s_1, ..., s_n] \), the OTCS compares two state-sequences \( S_m \) and \( S'_n \) based upon exact matching of the states (events) within \( S \) and \( S' \) (Zheng et al. 2010).

Although motivated by MESs, the OTCS does not utilize the hierarchical nature of MESs. Medical events use hierarchical coding standards such as the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Thus, the OTCS does not incorporate partial event matching, only counting the number of exact matches between event sequences. For example, if one MES contains ICD-9-CM code 250.00 and a second MES contains ICD-9-CM code 250.01, the original OTCS would not find a match although they represent highly related medical events. In addition, the OTCS does not allow weighting of matched events in its event matching component. For example, if two MESs share medical events 250.01 and 279.00, these matched events are given equal weight by the original OTCS. The original OTCS simply counts matched events, regardless of event likelihood, risk, or severity.

2.1.2 OTCS-MES with Prevalence Weights (OTCS-MES EP)

In contrast to the original OTCS, OTCS-MES integrates unique features of MESs. The OTCS-MES provides a matching component that integrates event prevalence, event duplication, and hierarchical coding, important elements of MESs. Event prevalence, normalized to mitigate heavy positive skew and compact distribution, provides weights for matched events. Partial matching captures similarity based on the hierarchical organization of event codes, increasing similarity beyond exact matching. For example, if one MES contains ICD-9-CM code 250.00 and a second MES contains ICD-9-CM code 250.01, the OTCS-MES matches these events at the 4-digit level but not at the 5-digit level (most specific ICD-9-CM codes).

Event prevalence weighting presumes that rarer events matched between two MESs indicate greater similarity than more common matched events. The OTCS-MES calculates individual event likelihood or prevalence using the complete set of trauma incident events and associated diagnosis codes. An event’s prevalence weight is one minus the event’s frequency rate, so larger values (weights) indicate rarer events. OTCS-MES normalizes the summation of prevalence weights of matched events by the maximum prevalence weight summation across all MES pairs. Additionally, OTCS-MES retains replicated matched events versus the original OTCS measure that removes replicated event matches.

2.1.3 OTCS-MES with Severity Weighting (OTCS-MES ES)

For trauma data, event severity provides intuitive appeal to weight matching events for mortality prediction. Early methods for mortality prediction incorporated injury scoring systems with event severity. Reference literature identified two factors most impactful in injury scoring, injury type and anatomical body region. Injury type describes the nature of the injury and includes values such as contusion, sprain, open wound, and dislocation. Body region involves the anatomical area of the body injured, such as head and neck, spine and back, torso, and extremities. Based on these two variables, Barell et al. (2002) developed a matrix having nature of injury columns, body region rows, and ICD-9-CM injury codes in each cell. As an extension to this work, Clark and Ahmad (2006) assigned a survivor proportion to each cell of the Barell matrix.

Our study uses the Clark/Ahmad extension with survivor proportions assigned to each ICD-9-CM injury code. A severity weight equals one minus the survivor proportion, with larger values indicating more severe events.

2.1.4 Trauma Mortality Prediction Model (TMPM)

TMPM (Glance et al. 2009), a probit regression model, uses approximately 1,000 different types of injuries characterized by coding sets such as ICD-9-CM. TMPM comprises two separate probit models. Model 1 uses all possible injuries as binary predictors with death as the binary outcome. Model 2 uses body region severity indicators. A weighted average of the coefficients of the two regression models provides the empirical injury severity for each injury.
Empirical analysis showed that TMPM ICD-9 provided superior performance than other ICD-9 based models. However, analysis in previous studies omitted analysis of operating points. Superior predictive ability of the TMPM-ICD-9 was most noted as the number of injuries increases (Cassidy et al. 2014).

2.2 Nearest Neighbor Classification for Mortality Prediction

The similarity measures described in Section 2.1 can be used in nearest neighbor classification algorithms. The kNN classification algorithm (Bhatia and Ashev, 2009) provides a simple but computationally intense approach for classification using a distance function. To make classification decisions, the kNN classification algorithm uses a neighborhood of $k$ nearest neighbors with majority voting among the $k$ neighbors. In this study, inverted similarity measures ($1 - \text{similarity}$) were used as distance measures.

To improve prediction performance, we use weighted voting and an ensemble of component nearest neighbor classifiers. Weighted voting allows more impact for neighbors close to a target case and less impact for far neighbors. The main benefit of weighted voted is less sensitivity to neighborhood size. We use proportional weights defined by Dudani (1976) as an alternative to traditional equal weighting voting. Ensembles combine predictions of individual classifiers typically using weighted voting among classifiers on each case. Ensembles improve classification results for diverse classifiers with different biases. Many ensemble methods have been proposed for nearest neighbor classification, using both training and voting to combine individual classifiers (García-Pedrajas and Ortiz-Boyer, 2009). We use a soft voting ensemble (scikit-learn.org/table/modules/ensemble.html) with cases labeled according to sum of predicted scores. This ensemble involves additional classification resources, as it requires determination of nearest neighbors for each component classifier.

With any classification algorithm, prediction of mortality in trauma data is a challenge due to imbalanced data. Despite the serious nature of patients admitted to trauma centers, mortality is uncommon. Treatment at a trauma center is short-term so only death between admittance and discharge counts as mortality. Patients dead on arrival and discharged to another facility do not count in the mortality disposition recorded in trauma data.

To deal with the uncommon mortality class, we used over sampling of the mortality class. Although Maloof (2003) indicates some conflicts in results of over versus under sampling, the availability of cases for the uncommon class drives usage of over sampling. Since ample data was available, we used the Over-Sampling Optimum Fraction (Kalton 1993) to increase the proportion of mortality events in trauma data from 6.28% to an optimum sampling fraction of 25.07% assuming equal data collection costs for both classes.

2.3 Research Questions and Hypotheses

This study asserts that a similarity measure adapted to medical event histories can be a valuable clinical decision-making tool. Within this broad assertion, this experiment addresses the predictive capability of MES-adapted similarity measures for the classification of trauma incident outcomes based on the incident’s set of events. We are interested in comparisons involving the predictive performance of classifiers using individual similarity measures (OTCS, OTCS-MES EP, and OTCS-MES ES), the existing standard for trauma morbidity prediction (TMPM), and an ensemble of nearest neighbor classifiers using individual similarity measures. We aim to observe improved prediction performance for OTCS-MES over TMPM and improved prediction ability of OTCS-MES relative to the original OTCS. The following list presents hypotheses concerning predictive performance.

1. TMPM, as the recognized best method, performs better than classifiers using MES similarity measures (OTCS, OTCS-MES EP, OTCS-MES ES, and OTCS ensemble).

As explained previously, TMPM has been designed specifically to predict mortality for trauma center incidents. According to Glance et al. (2009), since TMPM-ICD-9 performs better than other models, it should be preferred for risk-adjusting trauma outcomes when injuries are recorded using ICD-9-CM codes. Furthermore, Cassidy (2014) confirms the superiority of TMPM for injury scoring of adult patients especially as the number of injuries increases. Because trauma data standards mandate ICD coding, TMPM should continue as the preferred method for trauma incident prediction.
2. OTCS-MES performs better than the original OTCS similarity measure on morbidity prediction.

Unlike the original OTCS similarity measure, OTCS-MES allows generalized weighting of matched events and partial matching. These two capabilities should result in improved performance on trauma morbidity prediction. Despite these shortcomings, the OTCS may still identify the most important matching events to predict mortality in trauma patients. Lack of coding detail may negate the advantage of weighted, partial matching. Coding detail depends on data collection practices at trauma centers and perhaps beyond trauma centers with some ICD codes reported in a patient’s medical record before a trauma incident occurs. The original OTCS may match predictive performance of OTCS-MES with large amounts of cases and large neighborhood sizes.

3. OTCS-MES using event severity weighting (OTCS-MES ES) performs better than the OTCS-MES using prevalence weighting (OTCS-MES EP).

Injury severity is an appropriate weighting method for scoring trauma incidents based on referential literature. Also, injury severity has already been quantified by several scoring systems. Scoring systems based on ICD-9-CM codes, injury type, and anatomical region have been found effective in classification experiments. OTCS-MES, weighted by an event severity score (the Barell matrix survivor proportion), should demonstrate improved performance for trauma incident classification.

4. The best ensemble combining individual similarity classifiers should perform better than individual similarity-based classifiers.

Ensembles improve performance of diverse classifiers. We expect enough diversity between event matching based on exact matching, normalized event prevalence with partial matching, and event severity with partial matching to achieve improved prediction results.

These four hypotheses entail 10 statistical tests of classification performance. Tests 1a to 1d compare TMPM to OTCS, OTCS-MES EP, OTCS-MES ES, and the ensemble. Tests 2a and 2b compare OTCS to OTCS-MES EP and OTCS-MES ES. Test 3 compares OTCS-MES EP and OTCS-MES ES. Tests 4a to 4c compare the ensemble to OTCS, OTCS-MES EP, and OTCS-MES ES.

2.4 Trauma Data Set

We use the publicly available National Trauma Data Bank (NTDB) for our morbidity prediction experiment. Many hospitals contribute to the NTDB (www.facs.org/quality-programs/trauma/ntdb), a large aggregation of trauma data conforming to the National Trauma Data Standard. Specifically, we randomly selected test and training data from the complete set of 2015 trauma incidents in the collected NTDB trauma registries for 2015. Table 1 summarizes filters applied to the trauma data, compatible with filters used during development of the TMPM (Glance et al. 2009).

<table>
<thead>
<tr>
<th>Filter Description</th>
<th>2015 NTDB Trauma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data set</td>
<td>917,865</td>
</tr>
<tr>
<td>(1) Excluded incidents with all diagnoses being non-trauma (based on MARC table)</td>
<td>728,309</td>
</tr>
<tr>
<td>(2) Excluded incidents for patients w/age LT 1 year, or missing age or gender</td>
<td>685,587</td>
</tr>
<tr>
<td>(3) Excluded incidents with missing discharge disposition (HOSPDISP n/a)</td>
<td>590,288</td>
</tr>
<tr>
<td>(4) Excluded incidents w/patient DOA or w/transfer to another facility</td>
<td>427,545</td>
</tr>
<tr>
<td>(5) Excluded incidents for facilities handling LT 500 incidents during the year</td>
<td>403,534</td>
</tr>
<tr>
<td>(6) Excluded incidents having fewer than 5 diagnosis (event) codes</td>
<td>175,319</td>
</tr>
<tr>
<td>(6a) Deceased Disposition (6.28%)</td>
<td>11,010</td>
</tr>
<tr>
<td>(6b) Non-Deceased Disposition (93.72%)</td>
<td>164,309</td>
</tr>
</tbody>
</table>

Table 1: Summary of Filtered Trauma Data

From the 175,319 incidents having at least five events, we randomly selected 50,000 trauma incidents for a case base and 2,000 cases for testing. The training data set contains 465,325 total diagnosis codes (4,053 unique ICD-9-CM codes). We used the same test set to evaluate all hypotheses. Due to a shortage of
deceased cases, the case base had a 22% deceased prevalence (versus the optimal 25.1%), yielding 10,900 deceased cases.

### 2.5 Performance Measures

For statistical evaluations, we use Area under the Receiver Operating Characteristic Curve (AUROC or AUC) as the primary performance measure. AUC provides a prevalence independent measure of discrimination ability in risk prediction models. AUC has several equivalent interpretations including the expectation that a uniformly drawn random positive example ranks higher than a uniformly drawn negative example. Calculation of AUC requires a ROC curve of classification scores. For nearest neighbor algorithms, we used voting proportions among nearest neighbors as classification scores.

We perform two-tailed tests of AUC using Mann-Whitney confidence intervals augmented with the Logit transformation (Qin and Hotolovac 2008). In a detailed simulation study (Kottas et al. 2014), the augmented Mann-Whitney intervals (MW LT) provided good AUC coverage, robustness to unbalanced sample sizes and normality departures, and reasonable power.

Although widely recognized as a measure of discrimination ability, AUC does not provide an operating point for a classifier. We evaluate operating points using the weighted Youden’s Index and the Neyman-Pearson criterion. Youden’s index (Youden 1950), computed as sensitivity + specificity – 1, ranges from -1 to 1. A value of 1 indicates a perfect test with no false positives or false negatives. Li et al. (2013) introduced the weighted Youden’s Index when sensitivity and specificity are not equally important. In trauma center operations, treatment options emphasize avoidance of false negative errors (predicted survival but death occurs) although costs are difficult to quantify. Thus, an operating point should prefer sensitivity. In contrast to the tradeoff in the Youden’s Index, the Neyman-Pearson criterion (Neyman and Pearson 1933) maximizes sensitivity at false positive constraint levels.

### 3. Results of Empirical Evaluation

This section presents results of the empirical evaluation covering hypotheses presented in Section 2.3 as well as additional analysis of operating points on a ROC curve. Results of the hypothesis testing are presented first as they are the major results.

#### 3.1 Hypothesis Testing Results

Table 2 presents MW LT confidence intervals and related p values addressing the primary research hypotheses. For Hypothesis 1, the results show effects for TMPM compared to each individual OTCS measure and the OTCS ensemble. For Hypothesis 2, the results show effects for both OTCS-MES EP and OTCS-MES ES versus OTCS. For Hypothesis 3, test results show effects between OTCS-MES EP and OTCS-MES ES at an alpha of 0.10. For Hypothesis 4, test results show effects for all three classifiers versus the ensemble classifier, demonstrating sufficient diversity among individual classifiers.

<table>
<thead>
<tr>
<th>Test</th>
<th>Classification Method 1</th>
<th>Classification Method 2</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>TMPM (AUC 0.8392 CI:0.8326-0.8458)</td>
<td>OTCS (AUC 0.7894 CI:0.7840-0.7948)</td>
<td>&lt; 0.0001 *</td>
</tr>
<tr>
<td>1b</td>
<td>OTCS-MES EP (AUC 0.8065 CI:0.8008-0.8122)</td>
<td>OTCS-MES EP (AUC 0.8065 CI:0.8008-0.8122)</td>
<td>&lt; 0.0001 *</td>
</tr>
<tr>
<td>1c</td>
<td>OTCS-MES ES (AUC 0.8194 CI:0.8120-0.8268)</td>
<td>OTCS-MES ES (AUC 0.8194 CI:0.8120-0.8268)</td>
<td>0.0056 *</td>
</tr>
<tr>
<td>1d</td>
<td>OTCS-MES Ensemble (AUC 0.8589 CI:0.8521-0.8657)</td>
<td>OTCS-MES Ensemble (AUC 0.8589 CI:0.8521-0.8657)</td>
<td>0.0037 *</td>
</tr>
<tr>
<td>2a</td>
<td>OTCS-MES EP (AUC 0.8065 CI:0.8008-0.8122)</td>
<td>OTCS (AUC 0.7894 CI:0.7840-0.7948)</td>
<td>0.0024 *</td>
</tr>
<tr>
<td>2b</td>
<td>OTCS-MES ES (AUC 0.8194 CI:0.8120-0.8268)</td>
<td>OTCS (AUC 0.7894 CI:0.7840-0.7948)</td>
<td>&lt; 0.0001 *</td>
</tr>
<tr>
<td>3</td>
<td>OTCS-MES ES (AUC 0.8194 CI:0.8120-0.8268)</td>
<td>OTCS-MES EP (AUC 0.8065 CI:0.8008-0.8122)</td>
<td>0.0524 **</td>
</tr>
</tbody>
</table>
Table 2: Statistical Testing Results for Hypotheses

<table>
<thead>
<tr>
<th></th>
<th>OTCS-MES Ensemble</th>
<th>OTCS</th>
<th>OTCS-MES EP</th>
<th>OTCS-MES ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>4a</td>
<td>(AUC 0.8589 CI:0.8521-0.8657)</td>
<td>(AUC 0.7894 CI:0.7840-0.7948)</td>
<td>&lt; 0.0001 *</td>
<td></td>
</tr>
<tr>
<td>4b</td>
<td>OTCS-MES EP</td>
<td>(AUC 0.8065 CI:0.8008-0.8122)</td>
<td>&lt; 0.0001 *</td>
<td></td>
</tr>
<tr>
<td>4c</td>
<td>OTCS-MES ES</td>
<td>(AUC 0.8194 CI:0.8120-0.8268)</td>
<td>&lt; 0.0001 *</td>
<td></td>
</tr>
</tbody>
</table>

ROC curves provide a visual representation of performance differences. In Figure 1, the ROC curve for the ensemble dominates all other ROC curves except for two small intervals. The ROC curve for TMPM dominates the ROC curves for OTCS and OTCS-EP at false positive values below 0.40. For false positive values above 0.5, the ROC curves switch with the two similarity measures dominating TMPM. The ROC curves for TMPM and OTCS-MES ES cross in several areas with OTCS-MES showing a small advantage at low false positive values, but TMPM showing a small advantage at larger false positive values.

3.2 Analysis of Operating Points

To provide insight about choosing an operating point on a ROC curve, we examine results across score thresholds. The OTCS methods show an increasing advantage over TMPM as the sensitivity weight increases. Figure 2 shows weighted Youden Index values as the sensitivity weight increases from equal sensitivity/specificity (1/1) to high preference for sensitivity (10/1). The ensemble dominates TMPM at all weight levels. The individual OTCS approaches (OTCS, OTCS-MES EP, and OTCS-MES ES) dominate TMPM at sensitivity weights above 3/1. The performance of TMPM remains relatively flat over the range of sensitivity weights, while the four OTCS approaches increase in a linear manner with an increasing performance improvement over TMPM.

---

1 *: significant at traditional alpha of 0.05. **: significant at alpha of 0.10. The family-wise error rate (probability of making at least one Type I error) for simultaneous testing of 10 comparisons is 0.22.
The ensemble approach also shows advantages over TMPM using the Neyman-Pearson criteria. As shown in Figure 3, the ensemble provides higher sensitivity at false positive constraints below 0.6. At false positive constraints above 0.6, TMPM and the ensemble show similar sensitivity values with some crossing between the TMPM and ensemble graphs. For the individual OTCS methods, TMPM shows an advantage for false positive constraints above 0.5. For small false positive constraint levels, the individual OTCS methods show higher sensitivity values.

![Figure 2: Weighted Youden’s Index by Method and Cost Ratio (50,000 cases)](image)

![Figure 3: Maximum Sensitivity for False Positive Constraints (Neyman-Pearson criteria)](image)

### 3.3 Discussion

As summarized in Table 3, the results in Table 2 support all hypotheses except 1d between TMPM and the OTCS ensemble. Results demonstrate strong evidence to reject null hypotheses (equal AUC performance) except for less evidence to support Hypothesis 3 involving OTCS-MES ES and OTCS-MES EP. Similarity measures using partial matching and matching weights provide better performance than simple matching. Event severity based on domain knowledge of injuries provides improved predictive performance than prevalence without domain knowledge of injuries. The OTCS ensemble provides improved predictive performance demonstrating enough diversity in the base classifiers (OTCS, OTCS-MES EP, and OTCS-MES ES). TMPM, developed with a large set of data and specialized domain knowledge, provides improved predictive performance than nearest neighbor classifiers using similarity measures.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Result</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>TMPM &gt; OTCS</td>
<td>Strong evidence to confirm</td>
</tr>
<tr>
<td>1b</td>
<td>TMPM &gt; OTCS-MES EP</td>
<td>Strong evidence to confirm</td>
</tr>
<tr>
<td>1c</td>
<td>TMPM &gt; OTCS-MES EP</td>
<td>Strong evidence to confirm</td>
</tr>
<tr>
<td>1d</td>
<td>OTCS Ensemble &gt; TMPM</td>
<td>Strong evidence of opposite effect</td>
</tr>
<tr>
<td>2a</td>
<td>OTCS-MES EP &gt; OTCS</td>
<td>Strong evidence to confirm</td>
</tr>
</tbody>
</table>
2b  OTCS-MES ES > OTCS  Strong evidence to confirm
3  OTCS-MES ES > OTCS-MES EP  Some evidence to confirm
4a  OTCS Ensemble > OTCS  Strong evidence to confirm
4b  OTCS Ensemble > OTCS-MES EP  Strong evidence to confirm
4c  OTCS Ensemble > OTCS-MES ES  Strong evidence to confirm

Table 3: Summary of Findings on Hypotheses

The AUC values for Hypothesis 1d (TMPM versus OTCS ensemble) conflict with expected results. The performance difference is large enough to have confidence that the ensemble provides improved performance to the established TMPM. The ensemble method does not require any training as compared to extensive training with a large data set by TMPM. TMPM may also require periodic retraining to deal with concept drift. However, the ensemble requires more classification effort, combining nearest neighbor search of three component classifiers. Indexing may be necessary to mitigate the additional resource usage for three nearest neighbor searches.

The AUC for TMPM (0.839) falls below the reported value (0.88) in Glance et al. (2009). A possible explanation for TMPM’s smaller AUC value is concept drift in more recent trauma data (2015) used in this study than used in the original TMPM study (2002 to 2006).

The ensemble method provides better performance than TMPM using an operating point on a ROC curve. With equal weight on sensitivity and specificity, the ensemble provides a more credible score threshold (0.55) than TMPM (0.20) as well as a slightly larger, optimal Youden value (0.5541 versus 0.5296). However, sensitivity values at the optimal Youden value seem too low so higher weighting for sensitivity seems likely in practice. The nearest neighbor methods (individual and ensemble) provide linear improvements in the weighted Youden value as the cost ratio increases. The ensemble approach also shows advantages over TMPM using the Neyman-Pearson criteria at false positive constraints below 0.5. If a false positive constraint of 0.4 is feasible in practice, the ensemble should provide sufficient sensitivity (0.9545).

An important assertion in this study is the importance of partial, weighted matching using OTCS-MES versus the original OTCS. Concerning the OTCS-MES benefits for partial matching, Table 4 indicates that over 70% of the matched events between MESs are partial. Essentially, the original OTCS misses all partial matches with 3 and 4-digit partial matches containing valuable similarity information.

4. Conclusion

We extended a similarity measure for medical event sequences (MESs) and evaluated its classification performance for mortality prediction using a data set of trauma incidents. We generalized the event-matching component of the Optimal Temporal Common Subsequence for MESs (OTCS-MES) with user-defined weights. We compared the performance of nearest neighbor classification using MES similarity measures to the Trauma Mortality Prediction Model (TMPM), an accepted regression model for mortality prediction for trauma patients. The comparisons used a substantial data set from the National Trauma Data Bank. The results demonstrated an advantage on ROC curves, AUC, and operating points for the ensemble of nearest neighbor classifiers.
We plan future work on classification performance of linked patient records and a query architecture for medical event sequences. Linked patient records combine patient characteristics and medical event sequences. We plan to extend mortality prediction combining medical events with key characteristics of trauma patients. We also plan to predict high-risk patients using patient characteristics and MESs containing both medical events and temporal structure. In a second area of study, we will develop a query architecture supporting both similarity measures for linked MESs and regular expression matching to capture important patterns in MESs. To evaluate the query architecture, we will cooperate with medical professionals and analysts to determine use cases and evaluate utility of query results in decision-making. We will also develop storage and optimization techniques for large databases of linked MESs.

Acknowledgements

The Department of Education GAANN Program (grant P200A150283), focused on Big Data Science and Engineering, provided partial support for this research through a fellowship.

References


Maloof, M. A. (2003, August). Learning when data sets are imbalanced and when costs are unequal and unknown. In ICML-2003 workshop on learning from imbalanced data sets II (Vol. 2, pp. 2-1).


