Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 4th, 2007

Strategies for Managing Missing or Incomplete Data in Biometric and Business Applications

Mark Ritzmann and Lars Weinrich

Seidenberg School of CSIS, Pace University, New York

Abstract: *As data volumes dramatically increase, so too does the amount of missing data found in any given database or set of records. In order for decision making, verification, and identifications systems to maximize their accuracy, strategies and methods must be in place to manage the missing data in a prudent manner. In this series of experiments, we applied fallback models, designed to account for insufficient samples sizes and missing data, to the data associated with a biometric-base keystroke recognition system. Tests were then conducted in order to determine if improvements in accuracy could be obtained and, if so, which type of model demonstrated the most improvement.*

1.0 Introduction

Data volumes are exploding. The University of California at Berkley, which has been tracking the production of digital information, estimated the amount of digital data the world produced doubles approximately every two years.

Dependence upon data is also exploding as more and more corporations, institutions, and individuals try to use data as an asset. Witness the rise in analytic application, on line analytical processing, data based expert system, sense and respond systems, and systems based on data mining.

Often overlooked or, at least, badly managed is the process or strategy for handling missing data. Missing data is unavoidable. And, as data volumes grow, so too does the volume of missing data. The way this missing data is handled (or mishandled) can have huge implications on the validity and certainty of any resulting analysis or conclusions drawn.

In this experiment, we use a biometric keystroke analysis database and system to test the effectiveness of various types of missing data strategies. The data is comprised of typing samples. Therefore the missing (or, in some cases, insufficient sample sizes that have the same effect as "missing") data is actually missing or unused letters in the typing sample.

The missing data strategies are tactically implemented through the of fallback models, a technique inspired by back-off models found in speech recognition systems. Initially three fallback models were developed and tested; a fourth model was then developed and tested based on an analysis of initial results. Each model is a hierarchy and when leaf samples were found to be insufficient, that element "fell back" to a superior node in the hierarchy – a node composed of that element plus like elements forming a composite. If that node also proved to be of insufficient size, it then went to a higher level and so on.

At the highest level, this study is also a commentary on and insight into heuristic methods vs. statistical methods – the bookends of missing data management strategies.

2.0 Overview of Missing Data Strategies

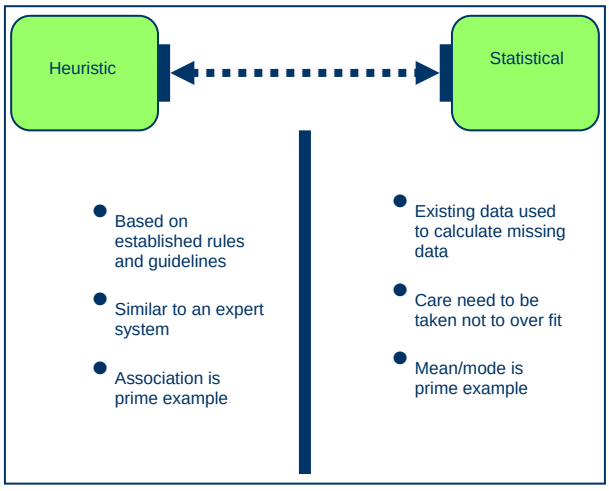
Generally speaking, missing data rates of less than 1% are considered trivial and seldom managed. Missing data rates of 1-5% are considered manageable. Missing data rates of 5-15% generally require sophisticated methods to handle, and any missing data rate of greater than 15% is considered to have the ability to severely impact any interpretation of conclusion drawn from that data [1].

Missing Completely at Random (MCAR) is the highest level of randomness. There is no dependency between missing attributes at all [3]. For example, data could be missing because the equipment malfunctioned, the data could have been entered incorrectly, of the subject never made it to the collection place. "Another way to think of MCAR is to note in that case any piece of data is just as likely to be missing as any other piece of data" [2].

Missing at Random (MAR) is the case where the probability of missing data on any attribute does not depend on its own value, but rather relies on the values of other attributes [3]. For example, if a survey asks a participant to report their income and families with lower incomes were less likely to report income that families with higher incomes, this missing data is MAR. "Just because a variable is MAR does not mean you can just forget about the problem" [2]. There is a relationship there that should be taken into account.

Not Missing at Random (NMAR) is the case where the missing data depends on the values that are missing [3]. *"For example, if we are studying mental health and people who have been diagnosed as depressed are less likely than others to report their mental status, the data are not completely missing at random. Clearly, the mean mental status score for the available data will not be an unbiased estimate of the mean that we would have obtained with complete data. The same thing happens when people with low income are less likely to report their income on a data collection form. The only way to obtain an unbiased estimate of parameters is to model missingness. In other words we would need to write a model that accounts for the missing data*" [2].

There are two basic means in which missing data can be obtained through data association. The first is to determine the value(s) for the missing data heuristically, usually accomplished by relying upon a domain expert who can understand and describe the related data [4]. The second is to rely upon data mining to examine the data and build a statistically based model which identifies probable, valid associations.



In practice, a perfect example of the Heuristic model would be an experienced police detective who, based on experience, can fill in and predict the values for missing information or data about a crime scene or suspect based solely on experience. In a sense, this technique employs an expert system approach and its ultimate success is dependent upon the skill and experience of that expert.

Statistical Models, on the other hand, rely upon the data available to predict the value of missing data. It is generally reported with confidence levels and follows the same rule, guidelines and limitation associated with data mining techniques.

Within this overall framework, some common methods to handle missing data include.

- Case Deletion – in this method, only instances without missing data are used – all other cases are deleted. This method, also called Complete Case Analysis, is the easiest manner to treat missing data and it is believed that all static software uses this method
- Parameter Estimation – an expectation maximization algorithm is used which estimated the allowable parameters of the missing data. Accuracy is dependent upon data distribution.
- Mean/mode Imputation – missing data is simply replaced by the mean of all available data (if numeric) or by the mode (if nominal data).
- Assignment of all possible values of the attribute – in this method, missing data is replaced by a set of variables comprised of all possible values for the missing element. This method comes with very large calculation costs.
- Regression Imputation – This method recognizes independent and dependent variables and then builds regression function to predict missing data.
- Hot Deck Imputation and Cold Deck Imputation – Hot deck imputation uses some of the existing data in the data set to impute values for missing data. Cold deck uses data that is not found in the current data set to impute the values for the missing data.
- Multiple Imputation – in this treatment method, more than two imputation values are received by running an imputation technique more than once. A vector is then constructed and used to replace the missing data.
- K-Nearest Neighbor Imputation – this method looks for the nearest sample to the missing sample and uses that value to replace the missing data [3].

3.0 As Applied to Biometrics Keystroke Analysis

On a broad level, there are two types of biometrics – physiological and behavioral. Physiological biometrics, such as finger print and eye scan, do not changes over time (with rare exception) and are nearly impossible to mimic. Behavioral, such as gait analysis and this experiment can change over time and are easier to mimic.

Biometric research and application has its root in security conscience organizations such as the Central Intelligence Agency and the Department of Defense. Generally speaking, the objective of a biometric identification system is either authentication or verification, with authentication being the easier of the two. Authentication is the act of certifying that someone is in fact who they claim to be (out of a finite set) and is a yes/no problem. Identification is a 1 out of n problem, attempting to assign an ID to a subject.

Our system is behavioral in nature and we test along both identification and authentication lines. Our hypothesis is that fallback models will improve the accuracy of the system

4.0 Fallback Models

Fallback models in this experiment are similar to back-off models found in areas such as voice recognition. The basic idea is that when one has insufficient data sampling for a particular value, that element falls back along a prescribed hierarchy to take a derived value, calculated by it and by others on the data set like it. In this way, the value of an insufficient element is still partially taken into account by the system as opposed to being simply discarded. In doing so, the hope is that the accuracy of the system is improved.

In our effort, three fallback models were initially tested and accuracy rates compared. Each model had two facets- duration and transition. The duration measurements are the length of time a key is pressed and the transition measurements are the length of time taken between the pressings of keys.

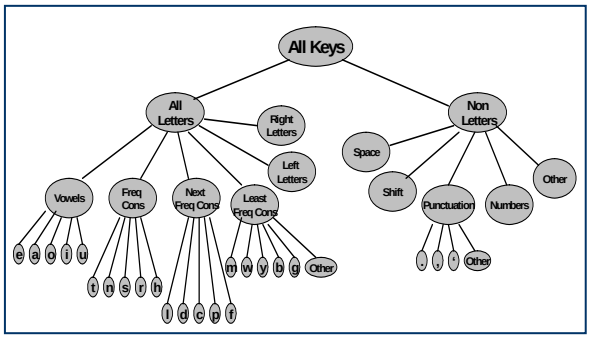
Later, after an initial round of tests and examination of those results, a fourth model was developed – the Hybrid model, which is detailed in Section 6.0

4.1 Linguistic Model

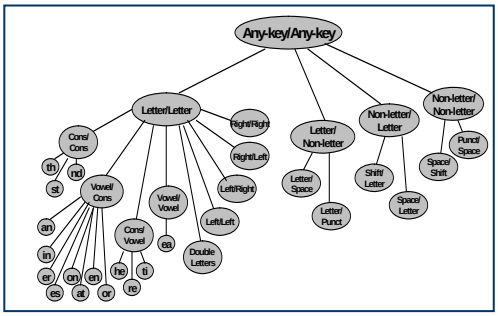
This model was developed in a study that occurred prior to this one. As such, the results of this model serve as a sort of baseline and a benchmark to be achieved in order to show improvement.

The Linguistic model is based of frequency – vowels (one of which appears in every word) are grouped together, as are most frequent consonants, next frequent consonants, etc.

A diagram on the duration aspect of the Linguistic model can be found below.



A diagram of the transition aspect of the Linguistic model can be found below.



4.2 Touch Type Model

To represent the heuristic end of the spectrum, a fallback model based on touch- typing rules and principles was developed.

The Touch Type approach was first introduced by inventor Frank Edgar McGurrin in the late 1800s. On July 25, 1888 he used this new technique to win a typing speed contest. His win spawned front page headlines across the country.

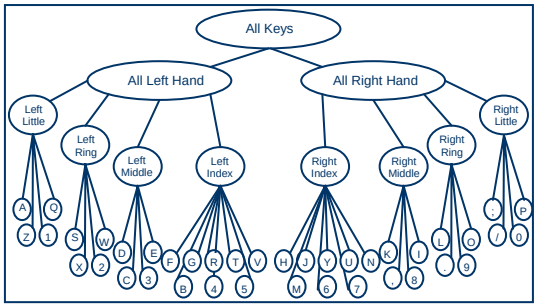
Touch type employs the sense of touch, rather than sight. Even today, most computer keyboards have a

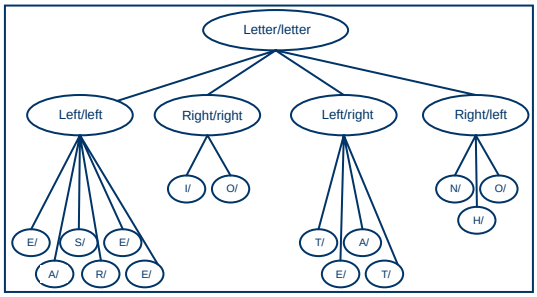
raised indicator on the “f” and “j” keys to signify the touch-type home position.

In this work, the touch type approach is viewed as the way to type and, therefore, adequately and rightly serves and as a heuristic based expert system

A diagram of touch-type finger-letter assignments can below.

The corresponding duration and transition models can be found below:





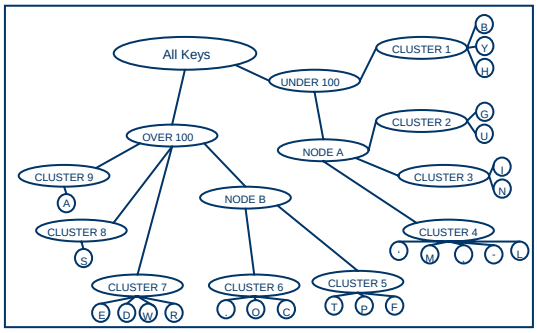
4.3 Statistical Model

To represent the opposite end of the missing data strategy spectrum, a purely statistical based model was also developed and tested. This model also encompassed duration and transition dimensions. The raw data from the 36 set was used, which consisted of the actual timings that were recorded.

To develop this element of the Statistical model, clusters were established on a near-neighbor basis. This clustering served to group leafs into appropriate nodes. Subsequent nodes, higher up on the model hierarchy, were then also derived based on closeness. The duration averages as well as the resulting clusters can be found in the figure below.

	Avg	Letter	Distance to Next
Cluster 1	93.65654603	B	0.313655407
	93.97020144	Y	0.574475253
	94.54467669	H	1.127778473
Cluster 2	95.67245616	G	0.314150302
	95.98660547	U	1.326165773
Cluster 3	97.31277124	I	0.348289117
	97.66106036	N	1.161750571
Cluster 4	98.82281093	Space	0.044389481
	98.86720041	L	0.146690579
	99.01389099	Coma	0.249303748
	99.26319473	M	0.276598746
	99.53989348	Single Quote	1.036537919
Cluster 5	100.5754314	T	0.660841455
	101.2362729	P	0.159288637
	101.3955615	F	1.902049416
Cluster 6	103.2976109	Period	0.107397967
	103.4050089	O	0.38375477
	103.7887636	C	2.051421506
Cluster 7	105.8401852	E	0.27896773
	106.1191529	D	0.51317892
	106.6323318	W	0.084624313
	106.7169561	R	6.073508362
Cluster 8	112.7904645	S	7.713464001
Cluster 9	120.5039285	A	

The resulting model can be found diagrammed in the figure below:



Regarding the transition dimension of the Statistical model, the 36 set was again used with the raw data consisting of the actual transition timings. However, prior to the development of this facet of the model, a treatment was applied. The purpose was to identify and remove outlier timings that reflected unnaturally long pauses between keystrokes. These pauses could be the result of the typist stopping to sip coffee, answer the phone, etc.

Outliers were then removed by the following process:

- The mean and standard deviation for each instance was calculated
- Timings that fell outside of the standard deviation were removed from the sample set; the sample set was therefore reduced

- The mean and standard deviation was then recalculated on the remaining data
- This process was repeated three times until the data and results stratified.

The overall data reduction can be found summarized in the figure below:

Data Compacting

	Sample Size	% of sample left after outlier wash
Total Round 0	11289	100
Total Round 1	8173	72.40%
Total Round 2	4982	44.13%
Total Round 3	2836	25.12%

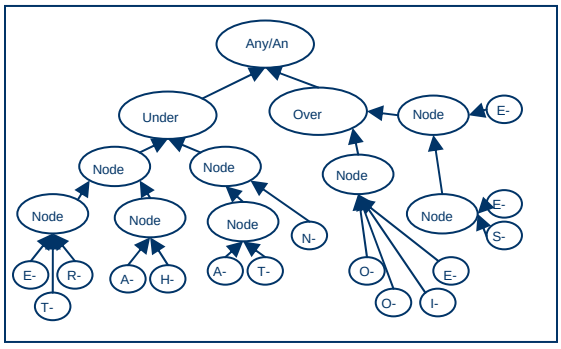
Data Compacting process

In essence, this technique equated to the application of a hot deck imputation. Approximately 25% of the total data was used to create a model designed to manage all of the data.

Similar to the means in which the duration aspects of the Statistical model were developed, the resulting averages were then clustered in a near-neighbor manner. Subsequent clusters, at higher level in the overall hierarchy, were then also derived. The derived averages and resulting clusters can be found in the figure below:

	Round 1	Distance to Next		Round 2	Distance to Next		Round 3	Distance to Next
E-R	23.01164478	2.01311438	E-R	21.16302	1.69486	E-R	20.91368	0.30915
			R-E	22.85788	0.43687813	R-E	21.22283	0.88851
T-H	25.02465916	0.03917741	T-H	23.29476	3.99832187			
R-E	25.06383657	4.89545905				T-H	22.11134	4.94556
			A-N	27.29308	2.60247			
H-E	29.95929562	0.53067801				A-N	27.0569	2.6899
A-N	30.48997363	6.77152742	H-E	29.89555	6.12301			
						H-E	29.7468	7.38735
A-T	37.26150105	5.69819861	A-T	36.01856	3.52108			
						A-T	37.13415	0.84697
T-I	42.95969966	2.9615527	T-I	39.53964	7.52213	T-I	37.98112	10.39564
N-D	45.92125236	8.39104418	N-D	47.06177	7.35842	N-D	48.37676	8.14744
O-R	54.31229654	6.23858077	O-R	54.42019	3.8416	O-R	56.5242	1.596
						O-N	58.1202	0.4717
E-N	60.55087731	3.00814565	E-N	58.26169	1.40621	E-N	58.5919	0.96942
			O-N	59.6679	2.01551	I-N	59.56132	12.34358
I-N	63.55902296	1.72425835						
O-N	65.28328131	4.10593409	I-N	61.68341	8.00324	E-A	71.9049	3.36792
E-A	69.3892154	7.10618775	E-A	69.68665	4.95049	S-T	75.27282	22.03619
S-T	76.49540315	17.86041491	S-T	74.63714	21.58324	E-S	97.30901	
E-S	94.35581806							
			E-S	96.22038				

The resulting model can be found diagrammed in the figure below:

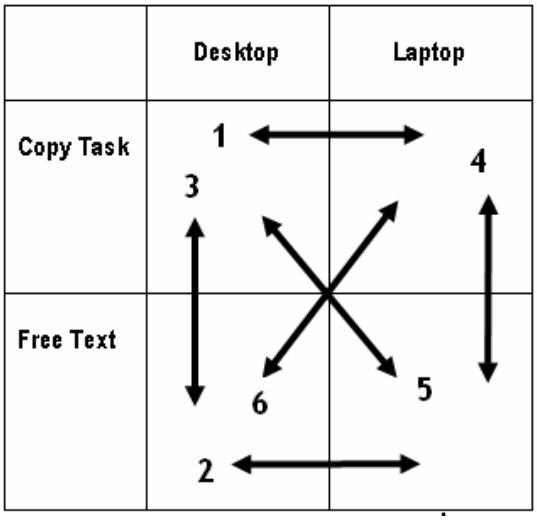


5.0 Experiments

The sample database was collected at Pace University in the fall of 2005. An online, java-based application was developed to facilitate the collection. Subjects were asked to provide sample using two different keyboard types – laptop and PC. Subjects were also asked to provide both copy samples from an established text and free typing samples.

Another application was then used to extract 239 features from each sample, forming the basis for sample comparison and, ultimately, subject identification.

Six different types of tests were then performed, as outlined by the diagram below:



There were two different tests conducted. The first was called "leave one out" where one sample per subject was removed, leaving four remaining samples per subject. The objective was to "replace" the removed sample back with correct subject.

The second test was called "train and test." This test was a straight-forward identification test with the goal of identifying each sample provider after the system was trained, via training data, accordingly.

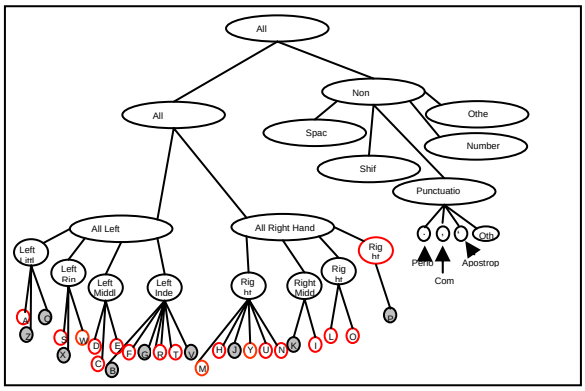
6.0 Results and Analysis

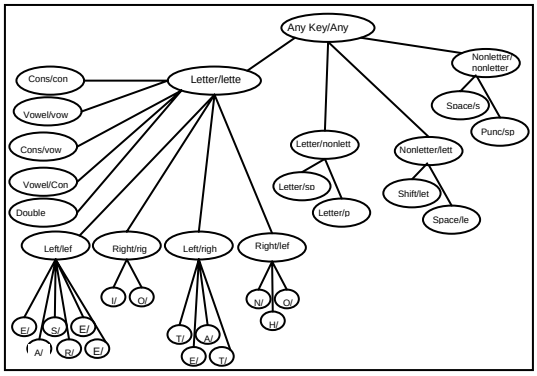
The results of the tests and comparison of the performance of the three models can be found below.

					Linguistic	TouchType	Statistical
Leave-One-Out	Copy	-	Desktop		99.45%	98.36%	96.17%
Leave-One-Out	Copy	-	Laptop		98.40%	98.40%	95.74%
Leave-One-Out	Copy	-	Combined		98.11%	98.38%	97.04%
Leave-One-Out	Free	-	Desktop		97.75%	95.51%	92.13%
Leave-One-Out	Free	-	Laptop		98.94%	96.81%	96.28%
Leave-One-Out	Free	-	Combined		97.54%	95.08%	92.98%
Test-And-Train	Copy	Desktop	Desktop		99.45%	98.36%	96.17%
Test-And-Train	Copy	Laptop	Laptop		98.40%	98.40%	95.74%
Test-And-Train	Copy	Combined	Combined		98.11%	98.38%	97.04%
Test-And-Train	Copy	Desktop	Laptop		63.83%	65.96%	56.38%
Test-And-Train	Copy	Laptop	Desktop		68.85%	69.95%	57.37%
Test-And-Train	Free	Desktop	Desktop		97.75%	95.51%	92.13%
Test-And-Train	Free	Laptop	Laptop		98.94%	96.81%	96.28%
Test-And-Train	Free	Combined	Combined		97.54%	95.08%	92.98%
Test-And-Train	Free	Desktop	Laptop		62.77%	60.11%	55.32%
Test-And-Train	Free	Laptop	Desktop		62.36%	62.92%	55.06%
Test-And-Train	-	Copy-Combined	Free-Desktop		82.02%	85.39%	68.54%
Test-And-Train	-	Copy-Combined	Free-Laptop		76.60%	77.66%	62.23%
Test-And-Train	-	Free-Combined	Free-Desktop		100.00%	100.00%	100.00%
Test-And-Train	-	Free-Combined	Free-Laptop		100.00%	100.00%	100.00%

The results indicate the Statistical model clearly performed worst. In addition, neither the Touch-type model or the Statistical model out-performed the Linguistic model, although the Touch-type model did show improvement is some test categories.

As a result of this information, a fourth model – the Hybrid model- was developed with the intention of combining the most effective elements of the Linguistic model with the most effective elements of the Touch-type model. The Hybrid model is diagrammed below.





The initial results of the Hybrid model (which are still being examined) seem to show this model did, in fact, perform the best.

7.0 Conclusions

As stated, this experiment is, at the highest level, a comparison between heuristic methods and statistical methods for handling missing data. While we will not claim to have shed tremendous insight into man vs. machine, our tests showed that the statistical model was the least effective by a fairly large margin. This came somewhat as a surprise.

Also surprising was the superior performance of the Linguistic model, given that is seemed to lack the logic and structure found in the other two initial models. The Linguistic model demonstrated the importance of the frequency element by, in essence, grouping together and focusing on the least frequent keys.

The Touch-type model (which could be argues is the most logical and heuristically pure of the set) showed some strength. In the end, combing the most effective characteristic from this model and with the frequency characteristics of the Linguistic model produced the best results.

References

[1] E. Acuna and C. Rodriguez, “The Treatment of Missing Data and the Effect on the Classifier” , Classification, Clustering, and Data Mining Applications. Spring-Verlag Berlin-Heodelberg publishers, pp 639-648, 2004

[2] D. Howell, “Treatment of Missing Data”, University of Vermont, posted at http://www.uvm.edu/~howell. Last revised March 23, 2007. Accessed April 9, 2007

[3] P. Liu and L.Lei, “Missing data treatments and NBI models” , IEEE Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications, August 2006

[4] C-H Wu, C-H Wen, H-J Chou, “Using Association Rules for Completing Missing Data” , IEEE Proceedings of the Fourth International Conference on Hybrid Intelligent Systems, February 2004.