Relationship Preserving Feature Selection for Unlabelled Time-Series

Fatih Altiparmak, Michael Gibas, Hakan Ferhatosmanoglu
Computer Science and Engineering
The Ohio State University
Columbus, OH 43210

Abstract

Feature selection has been widely studied in supervised datamining applications. In these supervised learning domains, the typical goal is to create effective natural clusters through the selection of a reduced attribute set that maximizes classification accuracies. However, an approach based on such a goal is not appropriate if the end goal is to preserve inter-attribute relationships,the data is unlabelled, and is in time-series format. All of these are typical characteristics of clinical trials data sets. In this paper, we introduce performance metrics and methods for feature selection over unlabelled time-series with the aim of preserving inter-attribute relationships. The proposed performance metrics estimate the effectiveness of a given feature set with respect to representation quality by measuring the nearest neighbors before and after feature selection. The proposed methods select features based on the time-series relationships between attributes through the incorporation of a technique that measures inter-attribute movements over the time-series. We provide techniques to combine and compare data from non-standard variable-length time-series sources and provide a mechanism to inject expert opinion into the feature selection process. The methodologies and comparative results are presented in the context of a real pharmaceutical database application.

Keywords: Feature Selection, Clinical Trials Data, Multi Variate Time Series Data

1 Introduction

Feature selection is the problem of selecting a subset of attributes based on an end goal. It is typically performed for the classification task where the problem reduces to finding a subset of attributes that categorizes the given objects as well as the full attribute set. However, little work has been performed for unlabelled data. As there are no class labels that assign similarity between objects exist in such data sets, the targeted task for this type of data is typically clustering. Each cluster is then accepted as a class label. This work introduces techniques to perform feature selection for unlabelled multivariate time-series data where the end goal is preserving the relationships between objects. The study is performed in the context of real clinical trials data and additionally provides techniques targeted to handle issues that arise with these, and many other real-world data sets.

Feature selection is a particularly important endeavor in pharmaceutical clinical trial studies. In such studies, an analyte is a measurable substance in the blood or urine such as hemoglobin, calcium or phosphate. Each clinical trial has a set of required analytes that need to be observed for every patient in the sample. This set varies according to the target disease of the study. Analytes are measured by various analytical chemistry methods where one or more chemical reactions produce some physically measurable quantity that has been shown to be highly correlated with the analyte concentration. In Vitro Diagnostic Devices(IVD) are hospital instruments used to measure analytes. Miniaturization of these devices is made possible with the recent technological advances in solid phase chemistry [1]. As a result, diagnostic testing is lead from out of central laboratories which enables a steady testing environment at sites close to patients, such as nursing homes, alternate care centers, and patient's home [2].

The most commonly marketed home-use IVDs have been the ones used to diagnose non-serious conditions (e.g. pregnancy) or to monitor conditions after diagnosis (e.g. diabetes monitoring). In the latter case, home monitoring is a continuum of the treatment [3]. A potential application of these devices is drug safety monitoring at home. A patient taking a drug with potential side effects could use such a device to get real time feedback on the effect of the drug on their health state. The patient could use this real-time knowledge in order to know if additional medical guidance is needed or if the drug should be discontinued. To this end, the relationship between analytes are important to monitor patient status.

Data interpretability and device capacity are key reasons to limit the number of analytes measured by a home-use IVD. In order for such devices to be economically feasible for home use, they will need to be affordable, and thus, limited in the number of analytes they can measure. Analyte measurements should only be included in the device if they pass a cost/benefit test. Since such devices will often be used by non-technical personnel, interpretation of results is an issue [4]. As the number of observed analytes increases, interpretation of results becomes more difficult and complex. Only the most useful analytes should be measured to maximize the accuracy of interpretation of the results. As the required set of analytes varies for different diseases, disease dependent IVDs should be developed that monitor a minimum set of analytes that is explicitly meaningful for that target disease.

Determining the minimal set of analytes that is appropriate to represent a particular disease is a feature selection task. The two main components of a typical feature selection technique are (1) performance metric and (2) the stopping criteria. The performance metric is used to select the attributes and this metric is chosen to be appropriate for the end goal. The feature selection algorithms are developed to find the subset of features that optimize this metric. The stopping criteria determines the size of the selected set. As the end goal of our task is to find the attributes that preserve the relationships between the objects, i.e. patients, the performance metric is based on how the K Nearest Neighbors (KNN) of an object (which are determined by using all attributes) are preserved on average after the feature selection task. The stopping criteria is based on the expected change on the performance metric when adding a new attribute. For real world analyte selection tasks, the subset should be selected using a combination of medical expertise and data analysis.

In this paper we discuss the data analysis component of selecting sets of analytes appropriate to monitor the overall health of patients with a particular disease or condition. Since all patients in the study have the disease, there is no further classification task to perform. The task is to find a global panel of analytes that represents well the overall health state of persons with that condition. We first discuss the challenges associated with clinical trial data and then propose a novel approach for preprocessing information from the clinical trial databases to facilitate feature selection. The overall approach involves transforming variable-length, time-series data into a high-dimensional constant size feature vector, and determining selected features based on processing this set of feature vectors. Attributes are selected as features based on their relationships with other attributes and the existing selected features. Attributes that have high correlation with unselected attributes and low correlation with selected attributes are favored with respect to the feature selection process. We also discuss how to incorporate expert opinion into this feature selection process.

The primary contributions of this work are:

- A feature selection technique that considers the relationships between attributes when determining appropriate features
- A technique to transform and compare non-standard time-series data sets
- Metrics to estimate the representation quality of a feature set for unlabelled data
- An approach that combines the speed of filter feature selection techniques with the greater accuracy afforded by the feedback associated with wrapper techniques using lightweight 'measure of goodness' computations
- A mechanism by which to incorporate expert opinion into the feature selection process.

The rest of the paper is organized as follows: First we explain our data transformation method. We define our representation quality metric and describe feature selection techniques. We present results and findings of the proposed algorithms. These algorithms and techniques are applicable to other data domains with similar characteristics, either dealing with feature selection for unlabelled data or combining heterogeneous data.

2 Challenges Associated with Clinical Trial Data and Unifying the Information for Each Patient: Data Transformation

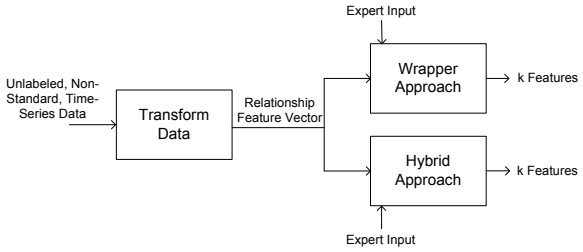

Figure 1: Overall Approach Architecture

Figure 1 shows a block diagram of our overall feature selection system. For our case, variable-length time-series data is transformed into a format that exposes important inter-attribute relationships and can be used to compare different data objects. After the data has been transformed, two approaches for feature selection, described in Section 4, were applied to select features such that the inter-attribute relationships of the selected features approximately represent all the inter-attribute relationships.

2.1 Challenges Associated with Clinical Trial Data There are a number of challenges in dealing with non-standard time-series clinical trial data [5]. The challenges addressed are the result of either:

1) Different intervals between measurements in the time series. Due to differences in patient/treatment availability or patient start conditions, the intervals between measurements can vary between patients.

2) Differences introduced by variability in data sources. These differences can be introduced because the trial is occurring at multiple sites, multiple clinical laboratories are participating, different evaluation techniques are used, or drugs are targeting patients with different sets of conditions.

3) Differences due to data mining or analysis techniques.

2.2 Transforming Time-Series Data for Analysis The proposed techniques address the characteristics of the data due to these challenges. The left portion at Figure 2 shows sample clinical trials patient data. For each patient, each analyte included in the trial is measured at various time intervals and a potentially different number of times. Data is transformed in order to eliminate the differences between the interval and the number of patient time-series readings. The patients participating in the same clinical trial have readings for the same set of analytes depending on the drug and the disease under the study. The number of relationships between analytes is the same for all of these patients. Differences resulting from different data sources or analysis techniques are handled by measuring the *relationships* between attributes rather than true attribute values. The goal of this particular task is not to find attributes and attribute values to classify patient state, but

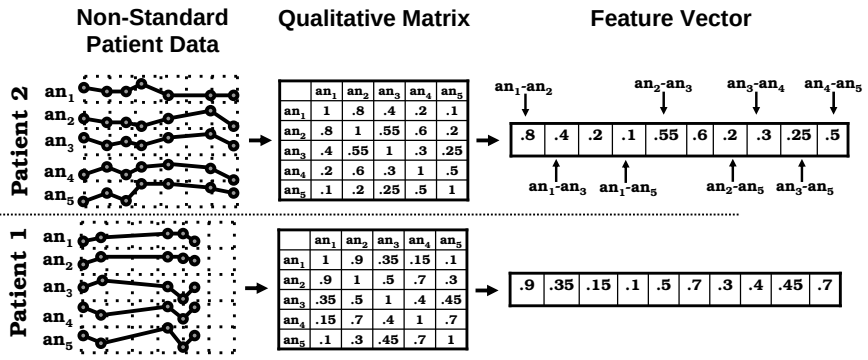

Figure 2: Patient data to Feature Vector

it is to find a global panel of analytes to effectively represent the health state of a patient having a given condition. It is necessary to generate a representation for each data object that contains the same amount of information in order to compare the objects.

Due to our particular problem domain, we selected a data transformation technique that measures relationships between attributes for each data object (patient). The inter-attribute distance metric is called the Qualitative Metric. It is meaningful for measuring inter-attribute distances of short time series and it reflects the way in which doctor's actually interpret patient data (comparing new readings to previous readings). Details about computation of the distance metric are provided in the next section.

2.3 Inter-Attribute Distance Metric: Qualitative Metric This metric [6, 5] compares movements between all possible (i, j) pairs which means, for a series of length n, it compares movements from first entry to each one of remaining $n-1$ entries, from second entry to each of remaining $n-2$ entries, and so on. The sum of the distances between all pairs is divided to the total number of pairs, which is $n*(n-1)/2$ and returned as the Qualitative distance.

The qualitative metric has been shown to be effective for short time series [6, 5]. It captures the similarity between patterns of changes of time series regardless of whether the nature of the dependence between them is linear or non-linear [6]. Qualitative distance uses a local distance function to compare the relationship between i^{th} and j^{th} entries of series s_1 to the relationship between same entries of series s_2. So, the local metric must be capable of comparing the relationship between the movements. We have proposed Slope-Wise Comparison (SWC) [5], to make this comparison. This local metric, SWC, takes four inputs: $s_1(i)$, $s_1(j)$, $s_2(i)$ and $s_2(j)$ and compares the relationship between $s_1(i)$, $s_1(j)$, and $s_2(i)$, $s_2(j)$. The method is called slope-wise comparison but instead of the duration, the sum of the absolute values of $s_1(i)$, and $s_1(j)$ is used to find an artificial slope. Let's call $AS1$ to the artificial slope between $s_1(i)$, and $s_1(j)$ and $AS2$ to the one between $s_2(i)$ and $s_2(j)$.

$$AS1 = \frac{s_1(i)-s_1(j)}{|s_1(i)|+|s_1(j)|} \text{ and } AS2 = \frac{s_2(i)-s_2(j)}{|s_2(i)|+|s_2(j)|}$$

These artificial slopes are compared to positive threshold (pt) and negative threshold (nt) in order to determine the distance between these two pairs. In case an artificial slope is less than 0, it is compared with nt. If it is less than nt then a value of -2 is assigned for that pair otherwise -1 is assigned. In the nonnegative case, if the slope is less than pt 1 is assigned otherwise 2 is assigned. The absolute value of the difference between the assigned numbers divided by 4 is assigned by the SWC metric as the difference between these pairs. Hence, there are five possible distances which can be assigned to a given time entry pair: 0, 0.25, 0.5, 0.75 and 1. The SWC metric checks if the ratio between two measurements of the same series are greater than a threshold or not. This is how the doctors compares the measurements of a patient. They compare current value of each analyte to their past values. Hence, we choice SWC as the inner metric. SWC returns if both analytes have similar behavior or not.

Because we are interested in inter-attribute relationships and can not directly compare our data objects, we transform each data object into a feature vector consisting of the inter-attribute relationship values. Figure 2 shows an example of raw clinical trial patient time-series data and how we transform it in order to perform data analysis. As is typical for clinical trials data, all analytes that are required to be measured for that trial are recorded at each observation point. The number of and intervals between observation points varies by patient. Although it is difficult to gain insights about relationships

between patients using analyte measurements, the analytes of the same patient can be compared to each other. The result of these comparisons can be represented as a matrix containing the qualitative metric computed for each analyte pair. A feature vector for the data object is generated by taking each unique pair of attributes from the matrix.

For our particular case, a clean subset of data was selected where each patient has at least 5 observations for each of 29 analytes that were measured as part of the clinical trial study. This data set consisted of time-series measurements for 857 Diabetic Neuropathy[1] patients for a industry sponsored clinical trial study. After data transformation, each patient has the same amount of information in their qualitative matrix.

Although we selected a particular metric to handle the heterogeneous, short time-series data, different transformations could be applied. For example, if the data has longer time-series and each data object has the same number of readings, a feature vector consisting of the correlation between each attribute pair could be used and may be more appropriate.

3 Feature Selection

In this paper, we define our task as: The goal of feature selection from n analytes is to find the subset of size k such that the relationships between these k analytes will best represent the relationships between n analytes. The two main components of feature selection are (1) the performance metric and (2) the stopping criteria. The typical feature selection performance metric is determined with respect to measuring classification accuracy for a given classification task. In our case, where there are no data categories, the metric must reflect representation quality rather than classification accuracy. The result set should contain the minimum number of analytes that represent the remaining well. Hence, setting the stopping criteria is an important issue.

3.1 Elements of Feature Selection

(1) The Inter-Attribute Relationship Performance Metric: As the performance metric is a measure for representation quality of the selected features with respect to all attributes, the selected metric should be a comparison between the representation using the original feature set and the selected feature set. We used the average of the starting index of the K nearest neighbors (K-NN) after removing the outliers ($SIKNO$) in this distribution.

We find the K-NN of a data object using all features and compare the position of the first occurring K-NN object using the selected features. K-NN of an object o using all features can be defined as:

$$KNN(o) = \{i | dist(o,i) \leq dist(o,j)\}, j \notin KNN(o)\},$$

where $|KNN(o)| = K$. The distance of each object to object o is then calculated by only using the features in the selected subset. The objects are sorted based on this distance in increasing order and the rank of each object in the $KNN(o)$ set is calculated. The starting index of the K-NN for an object o, $SIKNN(o)$ is defined as:

$$SIKNN(o) = min(Rank(i) | i \in KNN(o)).$$

We perform this for all objects and find an average value of the first occurring object among the nearest K neighbors over the entire set of data objects by

$$ASIKNN = \frac{\sum_{\forall o} SIKNN(o)}{TotalNumberofObjects}.$$

Intuitively, if an object is close to another object using both the selected features and all features, then the selected features provide a good representation of the distance relationships between data objects.

A variety of alternative metrics can be used that measure the preservation of distance relationships of the selected feature set. Outlier data objects can be removed before analysis or not. We can vary the K under analysis. We can also record the end index of the K^{th} nearest neighbor. We used the inter quartile range (IQR) based method to remove outliers from the distribution of the starting index of K-NN [7]. By discarding the effect of outliers on the resulting feature set, we find the set that is appropriate for the typical data objects in the dataset. Our performance metric, $SIKNO$, is calculated by the following formula.

$$SIKNO = \frac{\sum_{o|LL \leq SIKNN(o) \leq UL} SIKNN(o)}{|\{o|LL \leq SIKNN(o) \leq UL\}|}$$

[1] Diabetes is a life long disease marked by high levels of sugar in the blood. Diabetic Neuropathy is a common complication of diabetes in which nerves are damaged as a result of high blood sugar levels [9].

where LL, and UL are the lower and upper limits of the non-outliers set by the IQR based method. The denominator of this equation seeks for the number of objects having $SIKNN$ value within the normal range and the dividend is the sum of such values. In case LL and UL equals to the minimum and the maximum of the $SIKNN$ distribution, i.e. there is no outliers in the distribution, $SIKNO$ and $ASIKNN$ are equivalent to each other.

Because of the initial heterogeneity of our data set, we can not directly compare data objects against each other. We therefore measure similarity of patients as the Euclidean distance of their qualitative feature vectors. The feature vectors are in the unit hyper-cube, i.e. any value is within the $[0, 1]$ range. Hence, Euclidean distance is a natural choice between the qualitative feature vectors to give greater weight to greater distances within this range.

As an example, at the beginning there are 29 analytes, so the total number of analyte-analyte relationships is $28 * 29/2$, which is equal to 406. The K-NN is found using these 406 relationships. For the selected k features, the total number of relationships will be $(k-1)k/2$. For example, for $k = 6$ there will be 15 relationships. In such a case the new indexes of K-NN will be found by using these 15 relationships for each patient.

(2) Stopping Criteria: The stopping criteria determines the size of the selected subset. This chosen subset should meet the following two requirements: (1) The size of it should be as small as possible. Where as (2) addition of a new attribute should not lead a big change on the performance metric. There exists an analogy between selecting the number of features for this task and picking significant principle components for a factor analysis technique. Both should satisfy similar requirements. Our stopping criteria is inspired by the Kaiser Criterion [10]. This criteria picks the factors having Eigen values greater than 1. It is perhaps the best known and the most used criteria. In our case, if the expected decrease in the $SIKNO$ by an additional attribute is $\frac{1}{K}$, here K is the number of nearest neighbors used in $SIKNO$ calculations, then our feature selection algorithms terminate. For K equals 1, where $SIKNO$ is calculated based on the closest neighbor, our algorithms terminate at k attributes if the $SIKNO$ is less than or equal to $numberofattributes - k$ for the selected k features. For considering the closest K neighbors, they terminate when it is less than or equal to $\frac{numberofattributes-k}{K}$.

In the next section the algorithms will initially be introduced in a way that the size of the subset, k, is given. Then we will discuss how to perform feature selection when the subset size is not set initially.

3.2 Feature Selection Approaches Feature selection algorithms can be categorized as either filter or wrapper [12] approaches. The filter approach preselects the features according to some calculated 'goodness' measure. Whereas, the wrapper approaches incorporate the measure of goodness that has been used in an iterative feature search and selection [12, 11]. Feature selection for unlabelled data, is a relatively recent [11] research activity. The measure of goodness for both approaches for unlabelled data feature selection is based on the effectiveness of resultant clustering (i.e. maximum intra-cluster similarity and minimum inter-cluster similarity). Dy and Brodley [11] defined the goal of feature selection for unsupervised learning as finding the smallest feature subset that best uncovers "interesting natural" clusters according to given criteria in a static data set. In their technique, features are pruned if they do not add to the effectiveness of clustering metrics. Our situation differs in that we are not dealing with a ordinary time-series data set where each dimension has a single value for each object, but rather a multivariate time-series data set. Since we are interested in finding those analytes whose behavior is related within a data group, we instead focus on finding correlations between attributes. Because of the time required to compute clustering metrics, the existing work uses heuristics to prune some attributes based on some estimated measure of their effect on the clustering metrics. Our metric is not as computationally intensive, and we do not eliminate any attributes before analysis.

4 Proposed Feature Selection Algorithms

We explored two approaches, a forward-searching wrapper approach and a hybrid approach that combines a global feature-aware selection criteria and minimal performance computation. Our instance of the wrapper approach has been modified appropriately for the issues associated with handling attribute relationships. The performance metric $SIKNO$, which is a measure indicating distance preservation, is computed over the relationship feature vectors. For the hybrid approach, we combine an initial filtering step to seed a wrapper step that uses a much less computationally expensive measure of goodness that is not tuned to the end performance metric.

4.1 Wrapper Based Approach The idea behind the wrapper approach is to evaluate the selected feature set at each iteration of analysis and select the minimum set that leads to the most desired output for the feature set selection criteria. This approach was inspired by Kohavi and John [13] for feature selection in supervised learning. As the final goal is selecting k features out of n, we start with a subset of size k. This subset can be determined randomly, or by merging selected sets of size smaller than k. This discussion is left to the Experimental Results Section. We refine the set at each

Algorithm 1 :Wrapper Approach(k)

```
1   //k = number of features to select
2   selectedFeatures=Select k features randomly
3   minScore = SIKNO(selectedFeatures)
4   lastChange = k

5   for iteration 1 to ⌈k/2⌉
6       for i 1 to k
7           if lastChange == i
8             return selectedFeatures
9           currentFeature = selectedFeatures[i]
10          minForNi = infinity

11          for each feature_ni not in selectedFeatures
12              selectedFeatures[i] = feature_ni
13              curScore = SIKNO(selectedFeatures)
14              if curScore < minForNi
15                minForNi = curScore
16                feature_min = feature_ni

17          if minForNi < minScore
18            selectedFeatures[i] = feature_min
19            minScore = minForNi
20            lastChange = i
21          else
22            selectedFeatures[i] = currentFeature

23  return selectedFeatures
```

iteration by attempting to change one of the selected ones with an unselected one. To do this, the wrapper algorithm replaces that selected feature with each of the unselected features and calculates the $SIKNO$ score for the new set. If no unselected feature leads a lesser score than the feature in the set, the feature in the set is not updated and the algorithm continues with another feature in the set. If more than one feature yields a smaller score, then the one that results in the minimum score replaces the feature currently under analysis. The algorithm stops when we can not gain any improvement by replacing an attribute in the selected feature set. Algorithm 1 depicts the details of this algorithm. Primary differences between this adapted version and the original algorithm [13] are that (1) the performance metric in the original version is the classification accuracy where as it is $SIKNO$ in this version. (2) The original algorithm is designed for single point data and this version is devoted to multi-variate time series data. (3) Hence, replacing each attribute is equivalent to replacing $k-1$ relationships.

At each iteration of the outer loop, $SIKNO$ is calculated O($(n-k)k$) times. Hence, in the worst case, $SIKNO$ is calculated O($(n-k)k * \lceil k/2 \rceil$) times.

4.2 Light-Weight Hybrid Approach The final goal is to select the k features where the relationship between those features best represent the ones between the initial set of n features. A filter step is followed by a wrapper one in the proposed hybrid algorithm. In the filter step an initial set of k features is selected and this set is refined by the wrapper approach. The measure of goodness of the selected set, $SIKNO$, is calculated only after feature selection for comparative purposes. The criteria used in the wrapper step is not the computationally expensive $SIKNO$ as before, but r^2 as described in the following section.

r^2: A Measure of Goodness-of-Fit for Linear Regression The distance between data objects is calculated by the Euclidean distance which seeks for the closeness in the hyperspace. Assume that the relationship r_1 is selected but not the r_2. If by knowing r_1 we can predict r_2, then we can represent r_2 by r_1. This will decrease the difference between the Euclidean distances before and after the feature selection between two data objects.

The goal of linear regression is to adjust the values of slope, a, and intercept, b, to find the line that best predicts Y from X. This regression is formulated by $Y = aX + b$. The slope and the intercept are selected in a way that the sum of the square of the regression error is minimized. r^2, the square of the Pearson correlation coefficient [8], is the measure of goodness of fit for this prediction and it is calculated by:

$$\frac{[\sum_{i=1}^{m}(X_i-\mu_X)(Y_i-\mu_Y)]^2}{\sum_{i=1}^{m}(X_i-\mu_X)^2\sum_{i=1}^{m}(Y_i-\mu_Y)^2}.$$

Filter Step Initially the selected feature set is empty. In this step, the features are added one by one to the selected set. As the feature set of a patient no longer contains the readings of the features but the relationship between feature pairs, the first 2 features, f_1 and f_2, are inserted into the selected set. These two features are selected such that the sum of the square of r^2 metric between the relationship, $f_1 - f_2$ and others is maximum. This is summarized in Lines 6-14 of the algorithm shown in Algorithm 2.

Algorithm 2 :Filter Step(k)

```
1  //k= number of features to select
2  selectedFeatures = empty
3  f1, f2 //First selected Features
4  fnext //Next selected Feature
5  max_score = 0

6  for each relationship fi − fj
7      sum = 0
8      for each relationship fk − fl
9          sum = sum+(r^2(fi − fj, fk − fl))^2
10     if sum > max_score
11       max_score = sum
12       f1 = fi
13       f2 = fj
14 ADD f1 and f2 to selectedFeatures

15 for ai = 3 to k
16     max_score = 0
17     for each feature, ani, not in selectedFeatures
18         score = addCriteriaScore(fni)
19         if score > max_score
20           max_score = score
21           fnext = fni
22     ADD fnext to selectedFeatures

23 return selectedFeatures
```

The intuition behind adding new features is to maximize the correlation of a newly selected feature to the set of unselected features (i.e. maximize the representation of unselected features) while minimizing the correlation of the newly selected feature to the already selected features (avoid redundancy). As an example, when adding the 3^{rd} feature, f_3, a number of new relations represented by the added relationships are considered. The $f_1 - f_2$ relationship is already in a list of current relationships, and the new relations inserted are: $f_1 - f_3$ and $f_2 - f_3$. f_3 is selected such that the linear dependencies between these new relations and the already selected relationship, $f_1 - f_2$ are minimum and, that between unselected ones are maximum. For a given feature f_3, the $addCriteriaScore$ function, used in Line 17, first calculates the sum of the square of r^2 between any possible pair of the relationships $f_1 - f_2$, $f_1 - f_3$ and $f_2 - f_3$. This sum is named as the sum of the square of intra-dependency, $SSRD$. Then, the sum of the square of r^2 between new added relationships to unselected relationships are calculated and named as the sum of the square of inter-dependency, $SSED$. The desired f_3 is the one having the minimum $SSRD$ and the maximum $SSED$. Hence, $addCriteriaScore$ returns the ratio between these two sums: $SSED/SSRD$.

Wrapper Step At the end of the filter step, there are k features in the selected features set. The algorithm for the wrapper step is identical to the one shown in Algorithm 1, with 3 differences. These are, (1) the initial seeding of the feature set is derived from the filter step, (2) the metric used for the measure of goodness of the current feature set is computationally less expensive, and (3) we are maximizing rather than minimizing the metric. The features in the selected set are refined one by one. At each iteration of the inner loop starting at Line-6 (Algorithm 1), the ith feature in the selected set is replaced with

an unselected feature. This replacement is performed based on the $HybridCriteria$, the measure of goodness for this task. The details of this criteria can be found in Algorithm 3. The basic understanding behind it is that an unselected feature is represented by its nearest feature in the selected set. For a given feature set, this measure first finds all possible relationships between the features of this set. Then for each unselected relationship, the representor relationship is found in the selected ones. The sum of the r^2s between the unselected relationships and their representatives forms the value of the criteria. The wrapper algorithm tries to find the feature set having the maximum value for the $HybridCriteria$. While adding the new features, f_i, the measure utilized in the previous subsection, $addCriteriaScore$, aims to maximize the number of new relations represented by the added relationships. Hence, these two measures are similar to each other.

Algorithm 3 :Hybrid Criteria($selectedFeatures$)

```
1  selectedRelations = All pairs between selectedFeatures
2  Score = 0
3  for each relationship, rl, between the initial n features
4      max = 0
5      for each relationship, srl, in selectedRelations
6          if r^2(rl, srl) > max
7              max = r^2(rl, srl)
8      Score += max
9  return Score
```

The hybrid approach considers the correlations between the features and it tends to select the features that have the highest sum of correlation between the unselected ones. If the selected features can explain the unselected ones linearly, then there should be a correlation between the Euclidean distance between two objects using all attributes and that with using only the selected ones. Hence, our hypothesis is that the hybrid approach will return a similar value for the $SIKNO$ compared to the wrapper approach.

Both of these approaches have their own strengths and weaknesses. The wrapper approach will return a set having a smaller $SIKNO$ value then the hybrid approach for a given number of features, k. However, the running time of the hybrid approach will be much smaller than the wrapper one. Wrapper approaches can get stuck in a local minima hence, the result of the wrapper approach may change depending on the initial seed. Where as, due to the filter step, the hybrid approach always return the same set. On the other hand, both approaches can be modified to embed expert knowledge. An expert may identify a subset of the selected analytes, denoted by ea, and ask for the analytes that improve the representation quality of the other relationships given the expert-selected ones. For the wrapper approach, the first $|ea|$ analytes are set to these analytes. The remaining $k - |ea|$ are randomly selected and at each iteration of the outer loop of the algorithm in Algorithm 1 only the remaining are refined but not the first $|ea|$. Similar modifications should be applied to the hybrid approach.

In this section we introduced two main components, i.e. the performance metric, and the stopping criteria, that we supplied for the targeted feature selection task. We introduced two feature selection approaches build to optimize the performance metric for a given number of selected features, k. These approaches can be modified in a way that k is increasing one by one until the stopping criteria is met. The k is first set to 2 and at each iteration the obtained $SIKNO$ is compared to the stopping criteria. If it is greater than the value returned by the criteria, then k is increased by one and the new feature set is selected. For the wrapper approach, we can utilize the resulting sets of the smaller values while selecting the initial seed. Whereas, the filter step will only add a new analyte while increasing the k by one for the hybrid approach.

5 Experimental Results

We executed both techniques on our data set varying K to check the effect of K on the size of the resulting set returned by the stopping criteria. Recall that k is the number of attributes in the selected feature set and K is the size of the nearest neighbor set for which we are looking for the closest index in our performance metric $SIKNO$. The dataset contains 857 patients and for this set of experiments the wrapper approach is seeded randomly for each value of k. Results shown in Figure 3 depict that the two techniques yield nearly identical performance with respect to our inter-attribute relationship preservation metric $SIKNO$. However, the hybrid approach yields the same performance in much less time. In our case, because of the computationally intensive nature of the nearest neighbor computations, the wrapper approach takes several hours to compute a feature set. In contrast, the hybrid approach takes just a few seconds, an improvement on the order

of 3 orders of magnitude. The results also show that $SIKNO$ becomes much lower as we increase K. This result is obvious since we are increasing the size of the set of data objects that qualify as a nearest neighbor. We also see substantial improvements in the first few features we add, and very little improvement after a certain number, 8, and the algorithm stops adding new analytes at this point. Our stopping criteria leads our algorithms to return 8 analytes for the cases K equal to 1, 5, and 10 and they return 7 analytes for $K = 50$.

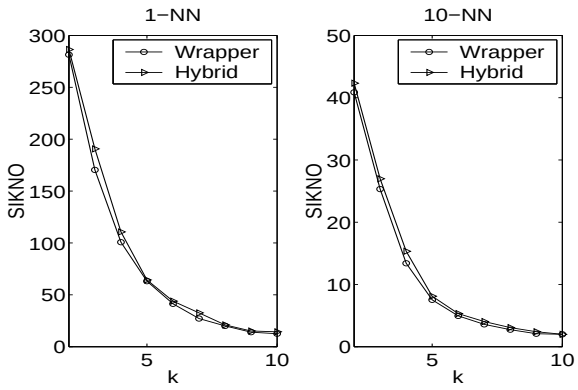

Figure 3: Relationship Preservation for proposed techniques as parameters vary

As mentioned at the end of Section 4, the wrapper approach may get stuck in a local minima. Hence the result set it returns depends on the initial seeds. For $K = 1$ and $k = 8$, we run the algorithm 100 times and we obtain only 8 different $SIKNO$ scores: 19.835, 20.079, 20.196, 20.212, 20.213, 20.352, 21.077, and 21.097. The greatest two are greater than the one hybrid approach returns which is 21.021.

We then performed this using smaller values of k and analyzed the relationships between the result sets for consecutive values of k by fixing $K = 1$. By doing this we expect to decrease the run time and to increase robustness of the wrapper approach. Generally the analytes in the result sets do not vary too much for different runs using the same value of k. For a pair of result sets from different runs using the same k, $k - 2$ of them are common and only 2 of them are different. In addition we were able to map each result set for runs using size k to a result set of size $k - 1$. In these mappings, at most, one or two of the analytes in the selected set are replaced by another analyte while increasing the size of the set by 1. For example, assume that the algorithm returns the set analyte$_1$, analyte$_2$, and analyte$_3$ for $k = 3$, then at most two of them are altered for the mapped resulted set in $k = 4$. Hence, for $k > 2$ seeding the wrapper approach with the result set from the $k - 1$ run and an additional random analyte from the unselected set seems reasonable. We ran the wrapper approach, for $K = 1$, applying this seeding strategy multiple times and each time we obtained the same result set having $SIKNO$ = 19.835. This is the overall minimum we previously achieved. Hence, the wrapper approach with the new strategy also stops at 8 analytes for $K = 1$.

We also compared the wrapper and the hybrid approaches against a technique where we randomly select a feature set 1000 times, compute the performance metric and select the best result. Both the hybrid and wrapper approaches yield better results than the best result out of 1000 random trials for $k = 8$. Table 1 shows the results for each of the approaches. The random and wrapper approaches take hours to complete while the hybrid approach finishes in a few seconds.

K	Wrapper	Hybrid	1000 Random
1	19.8	21.0	23.7
5	4.4	4.4	5.6
10	2.7	3.1	3.6

Table 1: K-NN Index for size 8 feature set

We also examined the effect of performing each step of the hybrid approach alone. Results were obtained from using only the result of the filter step and from applying the lightweight wrapper step to a randomly seeded initial feature set. As shown in Table 2, both the filter and wrapper steps contribute to the effectiveness of the hybrid approach. While each technique by itself performs well (they both yield results that are in the top 1% of the 1000 random feature selections),

executing the steps in sequence yields results that are better than the top result out of the random selections at minimal additional execution cost. Each technique runs in about a second by itself compared to hours for the 1000 random and wrapper approaches. The filter step helps the wrapper step avoid getting trapped in a local minima, while the lightweight wrapper step refines the feature set seeded by the filter step. When we combine the filter and lightweight wrapper steps, even though we are not using a feedback metric tuned to the end performance metric, we achieve nearly the same performance as when we do use the more computationally intensive metric.

It is important to note that although the size of our feature set is 8 out of 29, the number of relationships we are selecting to represent is $28/406$ ($< 7\%$). We achieve much better relationship representation quality than with using some arbitrary feature set. We can also obtain results very quickly using a metric that is not directly related to the end performance goal.

K	Wrapper	Hybrid	Filter Alone	Hybrid Wrapper Alone
1	19.8	21.0	36.7	31.7
5	4.4	4.4	6.7	4.6
10	2.7	3.1	3.8	3.1

Table 2: K-NN Index for size 8 feature set

Maintaining the Liver Hepatotoxicity Information Liver hepatotoxicity is determined by using the values of liver analytes. These analytes are Aspartate Transaminase (AST), Alanine Transaminase (ALT), and Alkaline Phosphate (ALP), Lactate Dehydrogenase (LDH), Gamma-Glutamyl Transpeptidase (GGT), Serum Albumin(SA), Total Bilirubin(TB), and Total Protein(TP).

Each of the rules used to determine the liver hepatotoxicity considers the values of a single liver analyte, hence there are named as single analyte rules. The followings are the rules considering the values of SA.

$$SA < 4.4 \text{ or } SA > 6.6 \text{ or}$$
$$SA.Day_k < SA.Day_0 * 0.8 \text{ or } SA.Day_k > SA.Day_0 * 1.2$$

Here, the first two rules imply that if the level of SA is less than 4.4 or greater than 6.6, which are 0.8 and 1.2 times upper limit of normal for SA, then the patient is accepted as hepatotoxic. The other 2 rules are comparing the SA levels under the influence of drug to the level of SA at the last day without the drug. Day 0 means the day just before the drug is induced. The days after that day is indexed based on their day difference to it.

There have been few attempts [14, 15] to change the rules for liver hepatotoxicity. For example "Hy's Rule" [14] requires the crossing of at least two of the 8 thresholds (TB and ALT are the ones examined in [14]). Otey et al. [15] proposed to apply Principle Component Analysis on the correlation matrix of the liver analytes for each patient. They used the first principle component as the feature vector, and defined the distance between two patients based on them. Their basic assumption is that the correlation between analytes are more important then the single analyte values for determining the liver hepatotoxicity. They implemented a basic outlier detection algorithm [16], KSUM, on top of these distances. This algorithm has 2 steps: In the first step the sum of the distances to the K closest objects is calculated for each object, and in the next step it sorts these sum of distances in descending order and declares the first couple of them as outliers. They computed these sums and checked the cumulative number of patients under the drug and placebo within the top portion of the sorted sums. They compared these values against the expected cumulative number of subjects for each rank. The basic assumption in this work is that if there is a hepatotoxic effect of the drug, then at top ranks the number of subjects on the drug should appear more than the expected. In Figure 4 we plot the cumulative number of subjects on drug and on placebo for the top 80 rankings by using all 29 analytes and the selected 8 analytes for $K = 10$ (This K is the $KSUM$ algorithm's input not the SIKNO's one). We used the result set of the hybrid approach for these experimentsby replacing AST with ALT as suggested by our expert[2]. ALT and AST are two most useful measures of muscle injury. Hence, these analytes are highly correlated. 100% of ALT is produced in liver, whereas AST has various sources, such as liver, skeletal muscle, and cardiac muscle. As ALT is a direct measure of the liver injury, the remaining experiments are done on the resulting set with this replacement.

The top two and the bottom two lines of Figure 4 express the cumulative number of patients on drug and on placebo given the ranking, and the middle two lines plot the expected cumulative number of subjects. The dataset we are using is a subset of Otey et al.'s D1 dataset. We got similar results, approved by an expert, as our colleges for the same dataset. As

[2]Expert

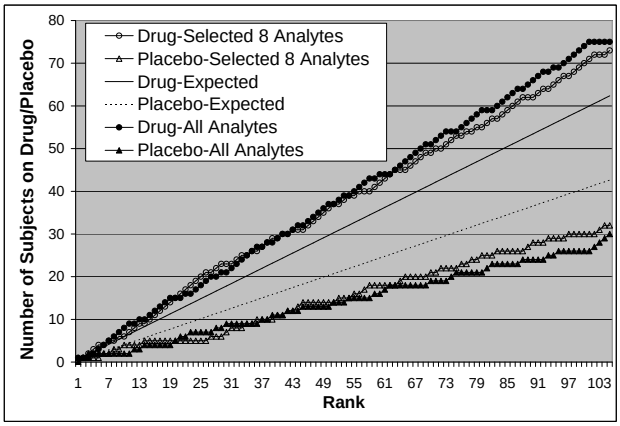

Figure 4: Outlier rankings for all analytes and the selected 8 analytes

can be seen from the figure, the hepatotoxicity content of all analytes is maintained by the selected 8 analytes using this technique. Additionally, the top 3 patients using all analytes appeared in the top 10 for the selected ones.

Another issue on liver hepatotoxicity for feature selection is to model the data to check if the hepatotoxicity information of the liver analytes is kept by the selected 8 analytes or not. The single analyte rules consider the minimum and the maximum of the values of the analytes and the minimum and the maximum of the ratio of these values under the drug to their values on the last day without the drug. Hence, for the selected 8 analytes we convert the data in a way that the result dataset has the following attributes: Drug, Sex, isHepatotoxic, Min_i, Max_i, MinRatio_i, MaxRatio_i. Here, Drug is a categorical variable having two values: drug, and placebo. Sex is the gender of the patient. isHepatotoxic is the binary class variable. Its value is determined by the single analyte rules. The term i gets values between 1 to 8 corresponding to the selected analytes. After fitting the data to a very simple logistic regression model where the correlations between drug and sex and other variables are considered, we get the following confusion matrix:

Actual ↓ / Predicted →	TRUE	FALSE
TRUE	47	19
FALSE	1	790

Females and males respond to the drugs differently and the drug does have a hepatotoxic effect hence this model is reasonable. As can be seen from the confusion matrix, even a simple model can explain this dataset with 71% accuracy for the hepatotoxic patients and 99.9% accuracy for the non-hepatotoxic ones. It is important to note that, while our techniques are designed to select a general purpose panel of analytes, the liver hepatotoxicity information maintained by the single analyte rules and the correlation based rule are also maintained by the selected set.

6 Related Work

This work is related to dimensionality reduction, as initially studied in statistics and existing feature selection research.

6.1 Statistical Methods Factor Analysis seeks to discover if the observed variables can be explained in terms of a much smaller number of variables called factors [18]. These factors are not necessarily the observed variables. The input of these analysis techniques is either correlation or covariance matrix between the variables. The main applications of them are (1) dimensionality reduction and (2) to detect structure in the relationships between variables [17] . The covariance matrix is decomposed into two portions, C and U. C is the common portion which is explained by the factors and U is the unexplained portion. Covariance matrix is the sum of these two matrices (U+C).

The aim of Multi-dimensional Scaling analysis is to detect meaningful underlying dimensions to explain observed similarities or distances between the investigated objects [19]. Any kind of similarity (dissimilarity) matrix can be analyzed using these techniques. In general, these techniques attempt to arrange the objects in a lower dimensional space in a way that preserves the order of observed distances between them. In other words, objects are arranged in lower dimensional space such that similar objects in the complete space are close to each other while dissimilar objects are far away from each other [20]. The problem we are targeting is similar. Our purpose of feature selection on the pharmaceutical data is to reduce the number of observed analytes in a way that preserves the NN relationship between inter-attribute relationships

for patients.

At the end of analysis, the underlying dimensions discovered by these statistical techniques are not the actual dimensions anymore. However, the resultant dimensions of these techniques are combinations of the observed ones. So, to find out the corresponding values of these new dimensions, one needs to observe most of the actual analytes. Our work differs in that we select features that are a direct subset of the initial feature set.

6.2 Feature Selection The feature selection problem can be divided into two categories based on the characteristics of the data they are applied on: supervised and unsupervised. In supervised feature selection [21, 22] , a feature set is selected such that the classification accuracy for the dataset improves or is maintained compared to the accuracy using all features. If there is a high correlation between the features a and b in the case of selecting a/b, selecting b/a is accepted as redundant. Method introduced by [22] take takes into account the correlations between attributes and does not add two highly correlated attributes to the feature set. In the filter step of our hybrid approach we try to find the set with minimum redundancy but for the unclassified data.

Yoon et al. [23] proposed a PCA based feature selection technique for multivariate time series data for classification. Their technique finds the principle components of each dataset separately and then combine the first p of them from each dataset for a particular class and find the descriptive ones for each class. These descriptive PCs got for each class separately are then clustered and the features contributing the most to the centroids of these clusters are taken as the selected features. As we are not getting the correlation matrix but the qualitative matrix, and our data is unlabelled, this technique is not applicable to our problem.

Unsupervised feature selection is a relatively new problem. The measure of goodness is the quality of the clusterings with the selected features. The aim is to select the features that discriminates the best. The basic assumption is that not all the dimensions at the beginning are informative. Whereas, we assume in this paper that all of them are. For more information we refer to [11]. Our work also differs in that we look to preserve relationships between objects and their aim is to find the features that minimize intra cluster dissimilarity and maximize inter cluster dissimilarity.

7 Conclusions

In this work, we have developed and applied a series of steps to perform feature selection that is targeted toward preserving the similarity of inter-attribute relationships between data objects. Heterogeneous time-series data with different numbers of readings for different data objects is transformed by generating feature vectors that capture measurements of relationships between pairs of attributes for each data object. This is performed in part to capture the inter-attribute relationships and in part to generate a homogeneous data set that allows comparison between data objects.

A performance metric is introduced so that inter-relationship distance before and after feature selection can be measured. Two Feature selection techniques are described. One is a wrapper approach that is tuned to this metric. Another, the hybrid approach targets inter-attribute correlation as the improvement measure of interest. A mechanism to inject expert input for feature selection is included so that known, important correlated inter-attribute relationships can be included in the selected feature set.

The hybrid technique, which uses a filter-type approach to arrive at an initial feature set, followed by a wrapper step using lightweight distance preserving estimate works nearly as well as the wrapper method which is tuned to utilize the specific measure of interest and works better than selecting the best feature set out of a large number of random permutations. The slight degradation in performance comes with substantial gains in execution time. We achieved about 3 orders of magnitude improvement using the hybrid approach.

Although the techniques included in this paper were developed for a specific problem domain involving unlabelled, heterogeneous, time-series data where inter-attribute relationships are more important for comparative monitoring than attribute values themselves, many of the ideas presented can be applied in a more general context. The proposed metric could be used in any feature selection application for unlabelled data, since the metrics are applicable to compare distances using any subset of features. The feature selection techniques can be easily modified in order to capture attribute value distances and correlations rather than relationship comparisons and correlations. Expert opinion can also be incorporated using the techniques provided in cases where certain attributes are desired in the selected feature set.

A recent data stream management systems (DSMS) design challenge focuses on maintaining some forms of Quality of Service (QoS) such as timeliness and data accuracy [24]. Most stream applications observe varying data arrival rates. Due to the vast number of continuous input data, i.e. streams, system resources (CPU, storage) may be consumed at highly variable rates, which leads to overloading and imminent degradation of QoS. Therefore recent DSMS designs and implementations employ sampling in order to overcome this issue [25, 26]. Within the data points arrived during the same time interval

sampling involves selecting a subset of them. The hybrid approach can be utilized as a sampling method in DSMS in cases where the relationships between streams should be preserved. This technique meets the requirements of the corresponding applications as it is fast and accurate.

8 Acknowledgements

The dataset and the funding of this project are provided by Pfizer, Inc. as part of a collaborative project about pathodynamics methodology.

References

[1] FDR Hobbs et al., *A review of near patient testing in primary care.*,Health Technology Assessment,(1997), 1:5.

[2] TH Grove, *In Vitro Dianostics: Bringing Testing to the Point of Care*,Medical Device and Diagnostic Industry (April 2000).

[3] Therapeutic Goods Administration, *IN Vitro Diagnostic Goods for Home-Use: Draft Guideline for Sponsors*, Australia Office of Devices, Blood and Tissues(June 2003).

[4] Medical Devices Agency, *Management and use of IVD Point of Care Test Devices*, DB 2002(03), London.

[5] F. Altiparmak, H. Ferhatosmanoglu, S. Erdal, DC Trost, *Information Mining over Heterogeneous and High Dimensional Time Series Data in Clinical Trials Databases*, IEEE Transactions on Information Technology in Biomedicine (TITB), Volume: 10, Issue: 2, pages: 254–263, April, 2006.

[6] L. Todorovski, B. Cestnik, and M. Kline, *Qualitative clustering of short time series: a case study of firms reputation data*, Information Society **A** (2002), 143–146.

[7] JW Tukey, *Exploratory data analysis*, Addison-Wesley Reading, MA, 1977:688pp. .

[8] MG Walker, *Pharmaceutical target identification by gene expression analysis*, Mini Reviews in Medicinal Chemistry **1** (2001), 197–205.

[9] A.D.A.M. Inc., *Diabetic Neuropathy(Nerve Demage -Diabetic)*,http://health.allrefer.com/health /diabetic-neuropathy-info.html, (May,2002)

[10] HF Kaiser,*The application of electronic computers to factor analysis*, Educational and Psychological Measurement, 1960, 20, 141–151.

[11] JG Dy, CE Brodley, *Feature Selection for Unsupervised Learning*, Journal of Machine Learning Research 5: 845-889, 2004.

[12] GH John, R. Kohavi, and K. Pfleger. *Irrelevant features and the subset selection problem*. Machine Learning, pages 121-129, New Brunswick, NJ, 1994. Morgan Kaufmann.

[13] R. Kohavi and GH John. *Wrappers for feature subset selection*, Artificial Intelligence, 97(1-2): 273-324, 1997.

[14] E. Bhornsson, R. Olsson, *Predicting the outcome of drug-induced liver disease*, Hepatology, 42(2):481–489, August 2005.

[15] ME Otey, S. Parthasarathy, DC Trost, *Dissimilarity Measures for Detecting Hepatotoxicity in Clinical Trial Data*, Proceedings of the Siam Conference on DataMining (SDM), 2006.

[16] EM Knorr, RT Ng, *A unified notion of outliers: Properties and computation*, In Proceedings of the thirth ACM SIGKDD international conference on Knowledge discovery and data mining(KDD),219–222, 1997.

[17] RB Darlington, *Factor Analysis*, www.psych.cornell. edu/Darlington/factor.htm.

[18] StatSoft Electronic textbook, *Principal Components and Factor Analysis*, http://www.statsoft.com/ textbook/stfacan.html.

[19] StatSoft Electronic textbook, *Multidimensional Scaling*, http://www.statsoft.com/textbook/stmulsca.html.

[20] SP Borgatti, *Multidimensional Scaling*, http://www.analytictech.com/borgatti/mds.htm, 1997.

[21] C. Ding and H. Peng, *Minimum redundancy feature selection from microarray gene expression data*, Proceedings of the Computational Systems Bioinformatics(CSB), 2003.

[22] L. Wu and C. Faloutsos, *Making every bit count: fast nonlinear axis scaling*, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining(KDD),664–669, 2002, Edmonton, Alberta, Canada.

[23] H. Yoon, K. Yang, and C. Shahabi, *Feature Subset Selection and Feature Ranking for Multivariate Time Series*, IEEE Transactions on Knowledge and Data Discovery, 17(9), 1186–1198, September 2005.

[24] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, *Models and issues in data stream systems*, Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS), 2002.

[25] N. Tatbul, U. Cetintemel, S. Zdonik, M. Chemiack, and M. Stonebraker, *Load Shedding in a Data Stream Manager*, Proceedings of the Very Large Data Bases (VLDB), Berlin, Germany, 2003.

[26] B. Babcock, M. Datar, and R. Motwani, *Load Shedding for Aggregation Queries over Data Streams*, Proceedings of the 20th International Conference on Data Engineering (ICDE), Washington, DC, USA, 2004.