Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2012, Article ID 539395, 11 pages
doi:10.1155/2012/539395

Research Article

Using Medical History Embedded in Biometrics Medical Card for User Identity Authentication: Data Representation by AVT Hierarchical Data Tree

Simon Fong and Yan Zhuang

Department of Computer and Information Science, University of Macau, Taipa, Macau

Correspondence should be addressed to Simon Fong, ccfong@umac.mo

Received 19 December 2011; Accepted 25 December 2011

Academic Editor: Sabah Mohammed

User authentication has been widely used by biometric applications that work on unique bodily features, such as fingerprints, retina scan, and palm vessels recognition. This paper proposes a novel concept of biometric authentication by exploiting a user's medical history. Although medical history may not be absolutely unique to every individual person, the chances of having two persons who share an exactly identical trail of medical and prognosis history are slim. Therefore, in addition to common biometric identification methods, medical history can be used as ingredients for generating Q&A challenges upon user authentication. This concept is motivated by a recent advancement on smart-card technology that future identity cards are able to carry patents' medical history like a mobile database. Privacy, however, may be a concern when medical history is used for authentication. Therefore in this paper, a new method is proposed for abstracting the medical data by using attribute value taxonomies, into a hierarchical data tree (h-Data). Questions can be abstracted to various level of resolution (hence sensitivity of private data) for use in the authentication process. The method is described and a case study is given in this paper.

1. Introduction

Biometrics has become increasingly common nowadays in authenticating users in security applications. There are many applications based on fingerprints, retina scans, voice waveforms, behavioural patterns and palm vessels recognition, and so forth. They work by the assumption that biometric resembles a bodily feature that uniquely belongs to an individual person and hardly anybody else. This biological feature is neither transferrable nor easily forged. A new kind of biometrics is devised in this paper, established on the information of one's medical history. Although medical history may not be absolutely unique to every individual person, it is very rare to have two persons who share exactly an identical trail of medical and prognosis. In fact, it is difficult to find any pair of persons who own exactly the same medical patterns in details that are described by time, location, age, diagnosis results, treatment dates and recovery progress, and so forth. It is therefore believed to be possible for using the pattern of medical history as a biometric in user authentication, at least in theory, in addition to the popular biometric identification methods. Similar biometrics theories are those based on one's email history patterns, online activity log patterns, and other personal history events [1, 2]. But medical history has its advantage because such history is relatively more difficult to be biologically forged, there are hard evidences that could be found from the wounds and scares; ultimate authentication by medical examination can be made possible for further verification, if necessary. The unique inerasable physiological feature favours biometrics authentication over other type of personal activities logs. The use of medical history can be implemented in a form of question-and-answer (Q&A) type of interactive challenge upon authentication, by supposing that only the authentic user has the secret (personal) knowledge about his or her past medical conditions. The information about one's medical

history can be a rich resource for generating Q&A challenges provided that the user has accumulated certain length of medical history.

This biometrics concept is motivated by a recent advancement on smart-card technology that future identity cards with gigabytes of in-built memory are able to carry patents' medical history like a mobile database [3]. Canadian airports are the pioneer that accept this kind of biometric security card for authentication and access control [4]; it is anticipated that many other countries and organizations will surely follow. The advantage of the original idea of embedding the medical history in a biometrics card is to allow medical rescue personnel access to this portable medical history from his card in case of emergency. Also the medical history on a card serves as a centralized depository because it could be handy when medical records are often stored in different hospitals. The history data stored in the card in principle shall be updated at the end of every visit to a clinic. With the full and latest medical records already in place in a portable biometrics card, ideally they offer a readily available resource for biometric authentication. Usually these medical records are stored in the memory chip of the card along with other popular digitized biometric data like fingerprint features too. Availability of the data is readily there on a portable biometric security card, what left of a research question is how these data could be used appropriately as biometrics for user authentication.

Two major challenges are projected here pertaining to using medical history as biometrics although the underlying archiving technology in a smart card can be safely assumed available. First is the process of matching and verification of lengthy medical history patterns in the task of authentication. Even it is technologically possible to store a longitudinal pattern of medical cases for a patient, obtaining a current pattern in the same longitudinal format (e.g., illness records from infant to current age) from a user as a test subject for testing or verification against his stored pattern during authentication task is almost impossible let alone accurate matching. If the testing pattern was to be acquired from oral interview with the user under authentication, it will surely be a very time consuming process. A quick method is needed for instant or almost instant authentication just like how prominent features of a thumb print are extracted from a scanned image in a very short time.

Sampling is one technique to tackle this problem when a full length of detailed data is not suitable for complete matching. More often, feature sampling which requires only a set of significant features to be matched has been used for biometric authentication [5]. Feature sampling is a general theme that includes using statistics, important events, and approximate outline of a series of events for instant authentication at a compromise of losing or omitting some details. Usually its efficacy is satisfactorily meeting some minimum performance expectation. Similar to feature sampling, sampling concept is to be applied on medical history data here, however, not by random; only some prominent features would be selected for authentication. This implies some mechanism is required for abstracting the medical history dataset into a lightweight representative pattern that can support efficient authentication. For example, a medical record that has specific attributes and values of the following: American, female, aged 19 months, suffered from meningitis, deaf and blind, would lead one to speculate she is Helen Keller.

The second challenge is privacy problem that is inherited from the nature of the medical history itself. Humans are generally uncomfortable to reveal too much detail of their private illnesses that show a sign of physiological weakness as a matter of ego. Since certain details of one's medical history are being taken as a personal secrecy for authentication, this secrecy would have to be confessed upon the authentication: the authenticator could be a machine or a human officer. Naturally this process of authentication operates in a form of exchanging simple questions and answers about the secrecy that the user holds, and it has to be fast and concise. The privacy challenge we face is to hide sensitive elements as much as possible in the message exchanges. In other words the questions would have to be asked implicitly without compromising the leak of the sensitive medical conditions.

If medical history was to be used as authentication data as an extra security measure, a special mechanism would be needed to protect the privacy of the data as well as an efficient data structure that can effectively hide and facilitate approximate matching of the medical patterns. Therefore in this paper, a new method is proposed for abstracting the medical data by using attribute value taxonomies (AVT), into a hierarchical data tree (h-Data). Questions can be abstracted to various levels of resolution (hence sensitivity of private data) for use in the authentication process. The method is described and a case study is given in the following section.

2. Proposed Solution

The solution for tackling the resolution of details regarding the medical history and privacy is to use h-Data by the transformation of AVT. Once the data are constructed in hierarchical format with the abstract data in a higher tier supported and related to the detailed data in a lower tier, questions can be derived selectively for user authentication. Figure 1 shows the process of converting a copy of the computerized patient's records into an h-Data that are stored together in a biometric smart card. The conversion process would be done at the level of certificate authority that can be trusted by users for data confidentiality. This paper focuses on how structured data with attributes in columns and instances in rows are converted to h-Data via aggregation and abstraction techniques.

After the h-Data are embedded in the biometric security card, it could be used for question-based authentication. Direct questioning can be done on the history data directly that is stored in structured format. Direct questioning is relatively simple because the questions can be randomly chosen from a set of facts from the structured table, and a binary verdict will return, should the answer matches or otherwise. Likewise, direct questioning can be done by simple visual inspection if the validator is a human officer,

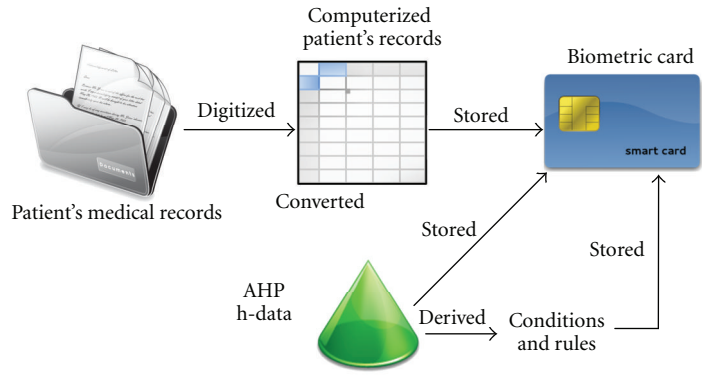

Figure 1: Conversion process of computerized patient's records to h-data.

for example the record shows a person has a limb amputated. Implicit questioning is a little more sophisticated that probes the user for answers that implicitly imply a medical condition. For example, for verifying if a patient is suffering from type II diabetes mellitus, implicit questions could be asking whether the user experiences hyperinsulinemia and obesity; asking the user questions about his daily diet in order to determine if he suffers from gastric disorders, or questioning his whereabouts in a specific period of time when his record shows that he was hospitalized, and so forth. Figure 2 shows the data stored in the biometric card can be used for two functions: computerized clinical records as recently proposed for convenience of medical consultant in different hospitals and for user identity authentication. In this case, the validator which is supposed to be a computer would be able to securely retrieve the h-data and from there derive a short list of questions to challenge the knowledge of the user with respect to his medical history. A rule checking module is necessary for cross-checking the answers from the users against the logics and the temporal orders of the facts in the h-data, for example certain medical conditions are likely to exist in a sequential order.

3. Representation of Medical History in AVT

Medical history data usually are comprised of various and meticulous clinical measurements, the data often carry many attributes. One of the challenges is to preserve privacy and find association among the attributes. In this paper, a multilevel data structure is proposed with the attributes flexibly abstracted and aggregated that represent various resolutions of the conditions of the illness. It helps hiding sensitive information by abstracting them and enabling checking in the form of Q&A with the testing user on the relations between the attributes of the data. We test the aggregation and abstraction techniques by using some sample data downloaded from UCI data repository (http://archive.ics.uci.edu/ml/) which is a popular site for providing data for benchmarking machine learning algorithms. The experimental results show that it is possible to appropriately abstract and aggregate medical data.

Many data prepreprocessing techniques such as data transformation, data reduction, and data discretization exist. However, these techniques are rather based on quantitative characteristics of the attribute values than the meanings of the attributes. Hence attributes are combined, transformed or omitted without referencing to their ontological meanings. For example, when these data are used in a decision tree that classifies heart diseases, the attribute that represents the number of blood vessels colored by fluoroscopy may get merged with another attribute that defines the number of cigarettes smoked per day, probably because they are just similar in mere numbers or statistical distributions as reflected from the prognosis data. Conceptually they may represent concepts from two totally different domains.

Apart from the broad spectrum of attributes and the depth of the associated values, another kind of complexity is the fact that the attributes and their values quite often are specified at different levels of resolution in a dataset. It implies that efficient methods for grouping and abstracting appropriate attributes are needed, while at the same time a consistent concept hierarchy or an organized view in relation to the multiresolutions of taxonomy must be maintained.

Attribute value taxonomies (AVT) that were proposed by Demel and Ecker [6] allow the use of a hierarchy of abstract attribute values in building classifiers. Each abstract value of an attribute corresponds to a set of primitive values of the corresponding attribute. However, the focus of the works in [7, 8] is formulating a new breed of learning classifiers, namely, AVT-decision tree that is hierarchical in nature for deriving rules directly from AVTs that are constructed from the data. This type of AVT-Decision is called h-data in this context here. For a simple example, the following diagram is a sample AVT that has a concept hierarchy of Season → phase of a season → month. The leave of AVT, that is, the month (June, July, August, etc.) can associate with abstracted attributes of a higher level. The abstracted attribute can in turn belong to that of a next higher level. If we have a set of decision trees, each is made for a different level or resolution in the concept hierarchy, we have the flexibility of testing or comparing cases that contain data represented in various resolutions.

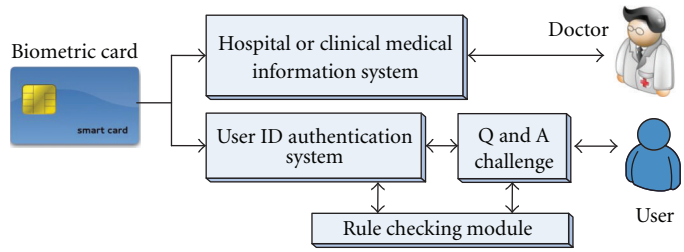

Figure 2: Workflow of the two uses of the h-data from the biometric card.

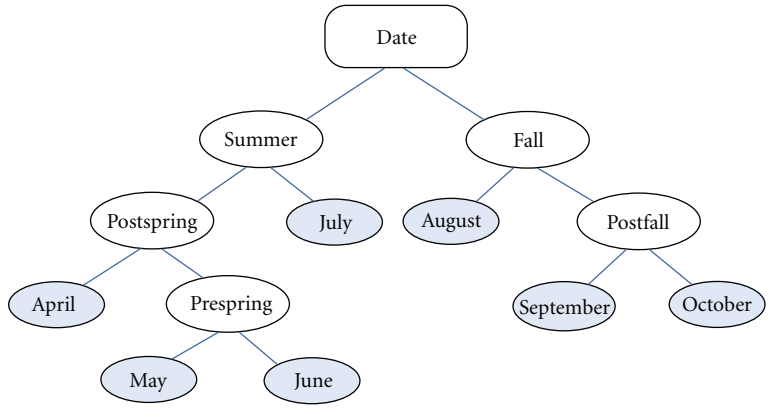

Figure 3: Sample AVT for date attribute of a dataset [6].

This approach is especially useful when we deal with data whose attributes have complex contextual resolutions. For clinical data records, a subset of attributes in the record may describe the body mass index (weight, height, plus even age, gender and race), another subset of attribute in the same record may represent the characteristic of a cell nucleus (radius, perimeter, area, smoothness, texture, etc.). The same goes forth for attributes that may describe other concepts in the context of clinical measurement, for example, insulin dose, (Regular, NPH, UltraLente dose). All these attribute may reside in a single record as a complete diagnosis. Some of the values and the units of these attributes may be the same, just like in Figure 3, but they belong to different concept groups, placed in different levels. Authenticators, however, are interested in knowing the interrelations among the attributes at different abstract levels, and in relation to the recorded decision, for deriving authentication questions. The decision tree which is represented by h-data serves as a hierarchical data structure that shows the causality (cause-and-effect) relations of the attribute data. The implicit questioning is based on principle of causality.

On the other hand, by generalizing and grouping attributes and their values to specific concept levels, the anonymity of the data can be enhanced, that satisfies one of the aims here for protecting one's privacy. Medical data are usually hierarchical. When the data are mapped into hierarchies, the specific data can become more general nodes in the hierarchy; hence the privacy can be better conserved. Sometimes some aspects of the data may be sufficient to identify a person especially rare illness.

In this paper, we devise a special hierarchical data model for allowing users to group data from a large set of attributes of heterogeneous natures, to organized concept views, similar to an AVT. The grouped attributes in abstract levels could be used for formulating questions during the authentication process in terms of how details the attributes are pertaining to a specific medical condition as the target class, and other interattributes relations. The challenge to be met in this model is grouping the attributes and then abstracting them to a higher level, which often requires expert knowledge or some common medical ontological databases. We used a collection of medical datasets as a case study, for evaluating the performance of the model.

4. Generation of Multilevel h-Data

The framework of the multilevel h-data generation model is shown in Figure 4. The central component in the framework is the preprocessing mechanism that receives two sets of data as inputs and transforms them into several datasets prior to decision tree building process. Decision tree is used here for knitting up the causality relations between the attributes, with a target class to which the model maps with the attribute data. For example, an illness of lung cancer would require inference from a number of smoking-related attributes, such as number of cigarette smoked per day, and years of being a smoker. The two input datasets are as follows, one is the original dataset with all the attributes, the other is a concept hierarchy represented in AVT format. The input of the concept hierarchy also specifies the number of levels

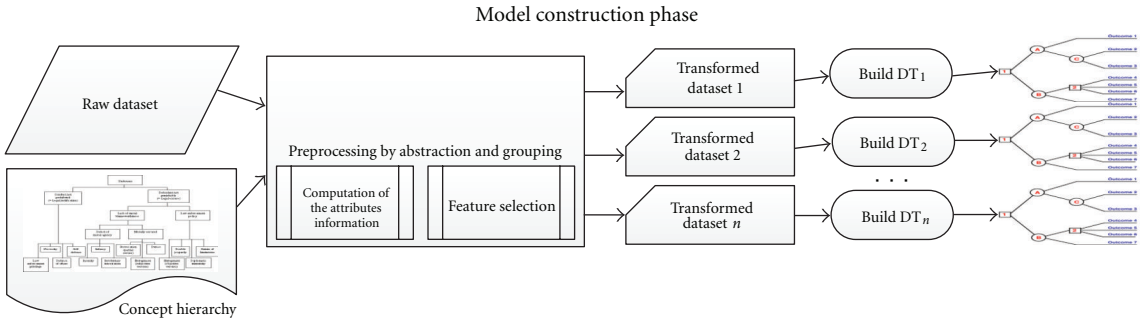

Figure 4: Framework of the Multilevel h-data generation process.

and what are the subgroups in each level. The concept of hierarchy is assumed to be defined by some domain experts such as medical doctors. The other input dataset is a full longitudinal history record of a particular person.

The output of the preprocessing is a set of transformed datasets that have been abstracted and aggregated according to their respective levels of abstraction at the concept hierarchy. There will be n number of transformed datasets (L_1, L_2, L_n), one dataset is for each layer of abstract concepts. The dimensions of the transformed datasets should be lowered down to the abstract concepts in the corresponding AVT level, such that $M = M_n \geq M_2 \geq M_1$, where M is the original dimension of the initial dataset, M_i is the new dimension of the transformed dataset L_i at level i. L_1 is the root of the AVT which also is the highest level, L_n is the dataset that has the M number of original attributes.

With the transformed datasets L_1 to L_n, traditional tree building process for example, C4.5 algorithm is used to induce the corresponding decision trees, DT to DT_n as outputs. Because of the reduced dimensionality the sizes of the trees follow this pattern: $C(\mathrm{DT}_1) \leq C(\mathrm{DT}_2) \cdots \leq C(\mathrm{DT}_n)$ where $C(\mathrm{DT})$ is the size of the DT in terms of the sum of nodes and leaves. Once the $\mathrm{DT}_{1,2\ldots n}$ are constructed they could be used for classification or prediction jobs by testing new data records. However, new data records now have the flexibility and options of taking any abstract form from whichever level of the concept hierarchy. The new data record needs to be transformed by the same preprocessing process (as in the model construction phase) unless it takes the same original dimensionality M as the original training dataset, prior to testing by the DT models.

The performance results as well as the information of the attributes during the model construction phase would be collected for visualization. With a large amount of description features, visualization in a hierarchy and groups of concepts offers easy comprehension to human readers of attributes information and the relations among them. One would be interested to know the general relations of two abstract concepts instead of the linkage of two detailed attributes. For an example of an authentication question based on medical history, whether and how much a seasonal climate that the user lives with or some general patterns of lifestyle that he is undertaking would contribute to his medical condition over time, make more sense, and are better interpretable than reading the measurements or very specific information on the individual attributes.

A compact decision tree that is built from abstract classes and attributes could potentially provide answers to high-level questions such as the example above.

Authenticators can try to find clues in the correct contextual level from the rules derived from such decision trees. And the questions can be derived from the relations of abstract concepts and their relations of prediction targets, instead of going to finer level attribute information, for formulating some general authentication questions.

The key mechanisms in the preprocessing process are the abstraction and aggregation methods. The two methods iterate from the lowest level to the highest up along the hierarchy specified in the given h-data according to the given concept hierarchy. The details of the two methods are discussed below. The overall operation of the model is depicted in pseudo code:

4.1. Aggregation Method. Aggregation is a common data transformation process in which information is gathered and expressed in a summary form, for purposes such as categorizing numeric data and reducing the dimensionality in data mining. Another common aggregation purpose is to acquire more information about particular groups based on specific variables such as age, profession, or income. Sometimes new variables would be created that represent the old ones while the new variables can better capture the meanings and the regularity of their data distributions.

We used two examples in our case study of organizing up some live medical data downloaded from UCI. One example is combining two attributes in the original data into a new attribute called body mass index that is more descriptive than the original ones. The two original attributes are *weight* (in kg) and *height* (in meters) to be put into a simple calculation. Sometimes categorical attributes are in text labels, crudely written; the language structures and grammars can be quite vague, depending on the sources. By using a lexical parser and analyzer, we analyze and rank the values of the multiple combined variables into a discrete measure of information completeness. New ordinal data may result, for example *highly contagious*, *contagious*, *neutral*; another example is *benign*, *malignant*, when specific formula is used to evaluate

```
Clean the data set from noise and missing values
Parse the ordered list of AVT and load them into memory
For i = level n to level 1.
Begin-For
    (1) Compute the attributes information in level i
    (2) Feature selection, eliminate redundant attributes if any: FS(D_i)
    (3) Aggregate selected attributes to abstract groups: Agg(D_i)
    (4) Abstract attributes to a higher level: Abs(D_i, L_i)
    (5) Consume the newly transformed dataset and build a corresponding decision tree: Classifier
        (L_i, DT_i)
    (6) Retain the performance evaluation results for visualization.
    (7) i--
End-For
```

Algorithm 1: Operation of the model.

the values across a number of the measurement attributes. The other example which is presented in Table 1, is on aggregating a set of conditional attributes that have binary values (true or false) into a single attribute. In the UCI medical dataset, there could be up to a dozen flags that describe the presence of a symptom, the seriousness of a symptom or the characteristics of a symptom. For example, in the heart disease dataset, combinations of conditional flags such as *painloc*: chest pain location (1 = substernal; 0 = none), *painexer* (1 = provoked by exertion; 0 = none), and *relrest* (1 = relieved after rest; 0 = none) are aggregated according to the abstract concepts in the AVT, into ordinal values of *high*, *medium_high*, *medium_low* and *low*. If the flags in each concept group are equally important, it would be a straightforward summarization by counting of true versus false. Or else, for the attributes carry unequal relative importance, the algorithm of multi-attribute decision analysis [9] is applied to estimate the ranks.

For the other attributes, categorical aggregation is applied based on the analysis of the number of distinct values per attribute in the data set. There are many ways of doing segmentation and discretization. Some typical methods include but not limited to binning, histogram analysis, clustering analysis, entropy-based discretization, segmentation and natural partitioning.

In our case study, a combined approach of binning and histogram analysis is adopted. The data are categorized by quartile analysis over a normal distribution of frequency. The quartiles (25% each) are used to grade the new ordinal variables as *low* $\leq Q1$, *medium_low* $\leq Q2$, and $>Q1$, *medium_high* $\leq Q3$ and $>Q2$, *high* $> Q3$. The aggregation applied here is unique from the traditional aggregation methods because the concept hierarchy structure is imposed by the AVT (predefined by experts). Two conditions must be enforced for transforming the data to be consistent with the given concept hierarchy. First the ranges and scales of the values associated across each attribute must be the same. Second, any new attribute emerged as a result of aggregating old attributes must be one of the concepts that exist in a next higher level up.

4.2. Abstraction Method. Abstraction here is referred to grouping attributes as guided by the AVT and systemically moving them on to higher level clusters in the tree hierarchy. If the full information on an AVT is available, it would be a matter of picking explicitly the attributes from a level and clustering them by aggregation according to a concept found in the next higher level. The process repeat until all the concepts are done, level by level in the AVT. The logical data format of h-data for representing an AVT would take the following form, similar to that in [10].

Let avt be an ordered set of subsets, where avt $\in$ AVT. An instance of AVT can take the following format:

$$\begin{aligned} avt\Big((number\ of\ concept,\ \langle concept\ names\rangle)_{level\ number}\Big) \\ = avt((1, \langle all\ diabetes\ records\rangle)_1, \\ (4, \langle insulin, glucose,\ exercises, diet\rangle)_2, \\ \ldots, (M_n\langle\cdots\rangle_n)), \end{aligned} \tag{1}$$

where M_i is the number of attributes, a, in level i, L_i is the working dataset in level i.

Dataset L can be viewed as a two-dimensional matrix such that $L_i = D_i(M_i,R_i)$, $i \in 1,\ldots,n$. Let $m_{\text{var}} = M_i$ and $r_{\text{var}} = R_i$, in level i. A dataset in D_i has m attributes that is, $D_i = (a_{1,i}, a_{2,i},\ldots, a_{m,i})$ with R_i instances in level i of *avt*.

As shown in the pseudocode in Algorithm 1, the function $\text{Abs}(D_i, L_i)$ is to partition attributes $a_{1,i}$ to $a_{m,i}$ from the original dataset D_i, in level i, and copy the new clusters of transformed data to level $i + 1$ in L_i. The purpose of the abstraction is to keep attributes in the same cluster to describe a common concept. The clusters themselves may be relatively different from each other. Therefore fewer clusters or concepts would be found in an upper higher level; the concepts are abstracted and can be described by using less attributes. For every i, except the root, L_{i-1} would contain a set of clusters to which the attribute a_i belongs. Such function is an optimization problem that uses heuristic to approximate solutions, if the information of the *avt* is not available, that is, we base solely on the information of the attributes and their values to form clusters. When the *avt*

Table 1: Examples of aggregation on binary variables in UCI medical datasets.

Lymphography data set	*M. Tuberculosis* genes data set
Lymphatics deformed?	ORF functions
Block of affere: no, yes	Class([1, 0, 0, 0], "Small-molecule metabolism")
Bl. of lymph. c: no, yes	Class([1, 1, 0, 0], "Degradation")
Bl. of lymph. s: no, yes	Class([1, 1, 1, 0], "Carbon compounds")
Breast Cancer Wisconsin (Prognostic) data set	*E. Coli* Genes data set
Cell symmetrical?	ORF functions
concavity	Class(5,1,1, "Colicin-related functions")
Fractal dimension	Class(5,1, "Laterally acquirred elements")
Smoothness	Class(5, "Extrachromosomal")

is fully available, the job is simply parsing the ordered lists and explicitly maps the attributes from D_i to L_i, attribute by attribute and level by level.

One of the abstraction methods, as studied by [11], is to measure the distances of the concepts and to determine how the concepts should be grouped by the attributes, should *avt* is not available even partially. It is called distance measures, which allows us to quantify the notion of similarity between two concepts. For an example of a medical record and assume somehow we have some missing information or uncertainty in a level of concepts in the *avt*, we may discover patterns from D_i such as "recovery duration is closer (more related) to age than it is to gender" based on distance measures. This kind of patterns presents ideas for grouping. If the similarity can be quantified, similar attributes can be quantitatively merged and labeled as a common concept.

Das et al. [12] proposed two approaches, namely internal-based and external-based measures to computing similarity metrics and they should be used together. Internal-based measure of a pair of attributes takes only into account of their respective columns, ignoring other attributes. External-based Measure is to view both attributes with respect to the other attributes as well. Distance is denoted as a distance measure function $d(a_i,a_j) = d(a_j,a_i)$ for attributes $a_i,a_j \in (a_1, a_2, \ldots, a_m)$. This measure maps the interattribute distance to real numbers.

Let v be defined as a subrelation over relation U that is written as $a_i = 1(U)$ where $a_i, a_j(a_1, a_2, \ldots, a_m) \in$. It is the enumeration of all tuples with attributes $a_i = 1$ or $a_i =$ true. Subrelation $v_{ai=1,\, aj=1}(U)$ is the enumeration of all tuples with $a_i = 1$ AND $a_j = 1$. The subrelations are denoted as $v_{ai}(U)$ and $v_{ai,aj}(U)$ for simplicity. Given a binary relation for U, two attributes are similar if their subrelations $v_{ai} = 1(U)$ and $v_{aj} = 1(U)$ are similar.

$$d\left(a_i, a_j\right) = \frac{v_{ai}(U) + v_{aj}(U) - 2 \times v_{ai,aj}(U)}{v_{ai}(U) + v_{aj}(U) - v_{ai,aj}(U)}. \tag{2}$$

Other possible implementations like those used in K-means are finding the similarity between two vectors of attributes, such as Euclidean distance, Minkowski distance, and Manhattan distance. It was already raised in [11] that the main problem is defining the right vectors and finding which attributes to constitute in it. So far it is still an open question

$$d\left(\overline{a}_i, \overline{a}_j\right) = \sqrt{\sum_{x=1}^{v} \left(\left| a_{ix} - a_{jx} \right| \right)^2}, \tag{3}$$

where a_i and a_j are vectors and v is the length of the ordered enumeration of the vector. For external-based measure, an extra working vector, E is needed and defined as $(e_1,e_2,\ldots,e_v)$ of size v. External-based measure is to compute the distance between a pair of attributes a_i and a_j with respect to E. One implementation proposed by Das et al. [12] is based on the marginal frequencies of the joint relation between a_i and each of the attributes in the external-set E.

$$d\left(a_i, a_j, E\right) = \sum_{e \in E} \left| \frac{v_{a_{i,p},e}(U)}{v_{a_i}(U)} - \frac{v_{a_{j,p}}(U)}{v_{a,j}(U)} \right|, \tag{4}$$

where $E \in (a_1, a_2, \ldots, a_m)$.

5. Experiment and Discussion

In order to verify the multilevel h-data model presented above, a number of data set were used in experiments to test out the outcomes. The medical datasets are obtained from UCI machine learning repository. It has been widely used by researchers as a primary source of machine learning data sets, and the impact of the archive was cited over 1000 times. The datasets used contain a relatively complex set of attributes with mix of numeric, Boolean, and nominal data types from various disciplines of biomedical applications. One of the clinical examples from the datasets used in our experiments is diabetics datasets provided by outpatient monitoring and management of insulin-dependent diabetes mellitus (IDDM). Patients with IDDM are insulin deficient. This can either be due to (a) low or absent production of insulin by the beta islet cells of the pancreas subsequent to an autoimmune attack or (b) insulin-resistance, typically associated with older age and obesity, which leads to a relative insulin-deficiency even though the insulin levels might be normal. Regardless of cause, the lack of adequate insulin effect has multiple metabolic effects. However, once a patient

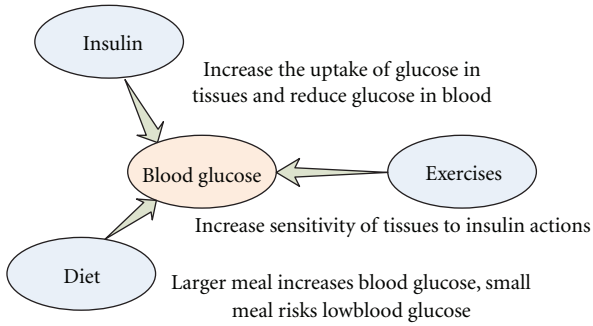

Figure 5: High-level relationship diagram of abstract groups.

is diagnosed and is receiving regularly scheduled exogenous (externally administered) insulin, the principal metabolic effect of concern is the potential for hyperglycaemia (high blood glucose).

Consequently, the goal of therapy for IDDM is to bring the average blood glucose as close to the normal range as possible. One important consideration is that due to the inevitable variation of blood glucose (BG) around the mean, a lower mean will result in a higher frequency of unpleasant and sometimes dangerous low BG levels. Therefore given the dataset which consists of a user's medical history records of his relevant diabetic's conditions, one record per clinical visit, an h-data model should be able to relate the blood glucose level based on the values of the other measurement attributes. We can see that the causality problem is somewhat complex because many attributes may contribute to the prediction target up to certain extent. And each of the interrelations of the attributes plays an influencing factor to the prediction. The last but never the least challenge is that the original attributes spread across different major concepts (insulin, blood glucose, body, and diet) and at different resolutions.

To tackle this causality problem, a multilevel h-data model is to be built. Firstly, we attempt to model an AVT on h-data that shows all the necessary concepts, at different level of resolutions/abstraction. We start by modeling the problem in the form of relationship diagram, as shown below. The relationship diagram in Algorithm 1 captures the essence of the main entities in the scenario. For simplicity, the attributes are yet to be shown. Combining the goal that is defined by three facets, with the main entities, we establish a conceptual hierarchy by attaching the corresponding attributes to them. Furthermore, between the lowest layer which has the original attributes and the level 1 of the hierarchy, several abstracted concepts have to be added in, by human judgments. The middle level forms an abstract view which would be used later in estimating the relations of the clustered attributes to the target class (which is one of the goals defined).

The target is defined by two objective, namely, abnormal blood glucose conditions and hypoglycaemic symptoms. The conditions are defined accordingly and they will be used to cross-check with the values of the respective attributes in the dataset. By doing this we establish a relation between a conceptual item (high blood glucose) with a number of refined measurements that often come in numeric. Conceptual items are useful for deriving authentication questions in biometric application because they can be relatively easier questioned and answered instead of numbers. (Who can remember a certain glucose test result in number on a specific date, e.g.?)

Abnormal blood glucose (BG) conditions are as follows:

(i) premeal BG falls out of ranges 80–120 mg/dL,

(ii) postmeal BG falls out of ranges 80–140 mg/dL,

(iii) 90% of all BG measurements > 200 mg/dL and that the average BG is over 150 mg/dL.

Hypoglycemic (low BG) symptoms are as follows:

(i) adrenergic symptoms, BG between 40–80 mg/dL;

(ii) neuroglycopenic symptoms, BG below 40 mg/dL.

Together with the full training dataset, the AVT would first be decoded in an ordered list format and fed into the preprocessing process as specified in Algorithm 1. The original attributes in the dataset would be aggregated and abstracted, as discussed above and transformed into a set of new datasets $(L_1, L_2, \ldots, L_n)$ ready to be consumed by the decision tree algorithms. Figures 5 and 6 demonstrate the results of the attributes being aggregated into four-standard categories. Some examples of attributes that are aggregated from continuous values to categories are shown in Figure 7.

The end result is the h-data which is a collection of decision trees with each specially prepared for the abstract concept views of a level in the AVT. An illustration is shown in Figure 8 where a cone shape which represents the h-data is in fact formed by a number of decision tree each of which shows the relations of attributes and groups and the groupings and hierarchy are predefined by AVT. By surfing up and down of the h-data, the authenticator can find the same but at different abstraction of information for formulating authentication questions. This is one requirement needs to be fulfilled for biometric authentication must be concise and fast. We illustrate the results here by building a visualization prototype that is programmed in Prefuse which is an open-source interactive information visualization toolkit and Java 2D graphics library. Through the selectors in the graphical user interface, we can have the options of choosing to view the combinations of the three domains of information.

(1) Predicted class: (center circle):

abnormal BG, premeal
abnormal BG, postmeal
abnormal BG, general
hypoglycemic, high BG
hypoglycemic, low BG.

(2) Link information: (line thickness):

predictive power to the target,
rank of relevance to the target,
information gain with respect to the target.

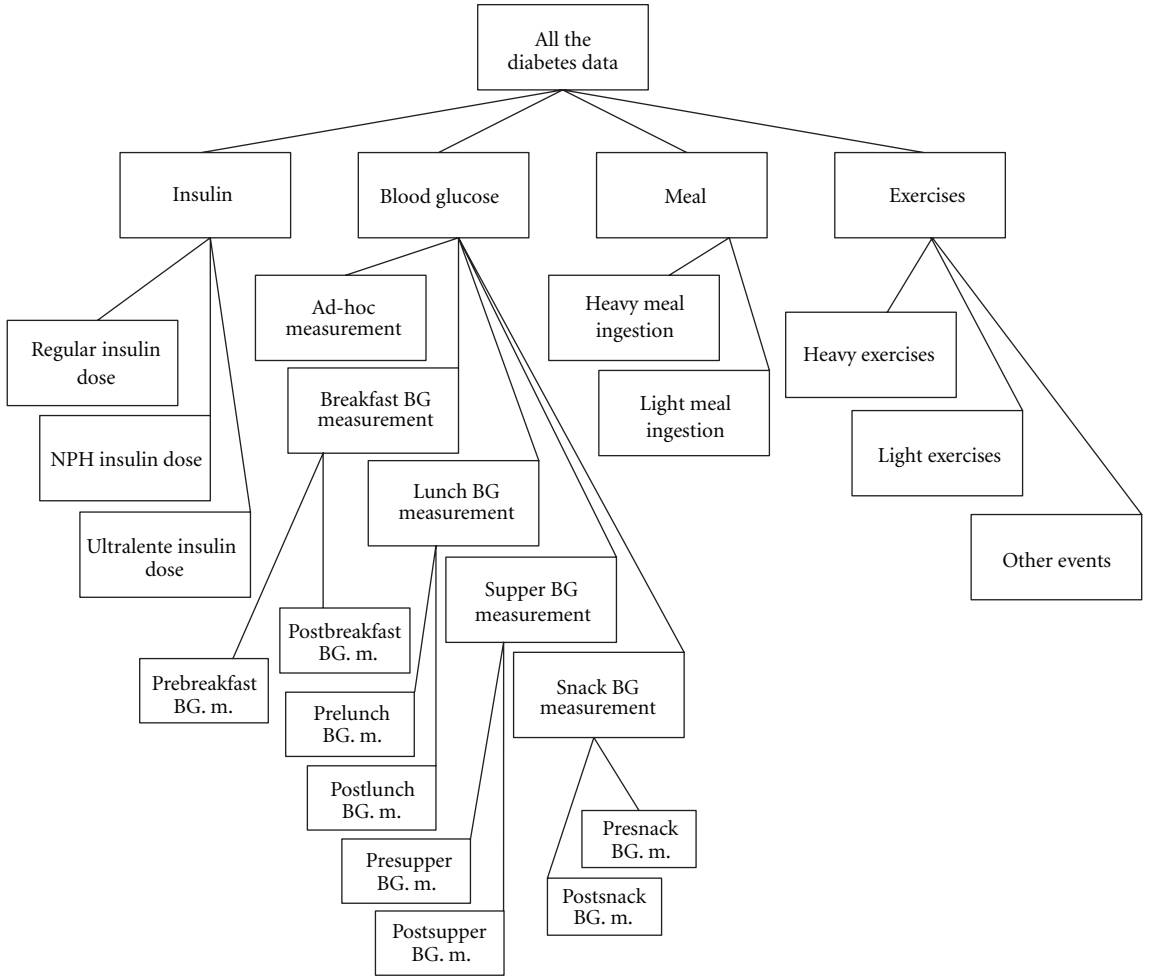

Figure 6: Concept hierarchy tree as AVT in the diabetics dataset.

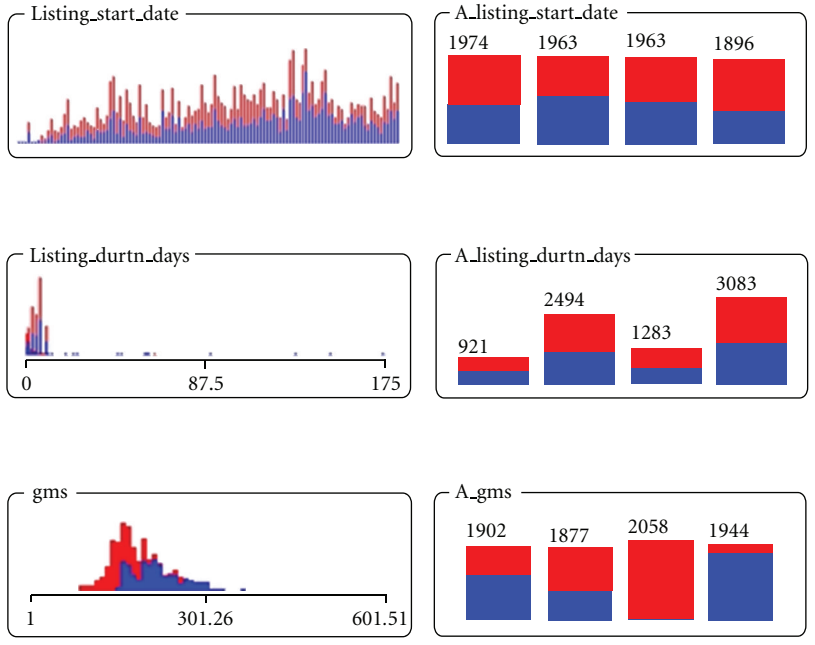

Figure 7: Snapshots of aggregated attributes (high, MedH, MedL, low).

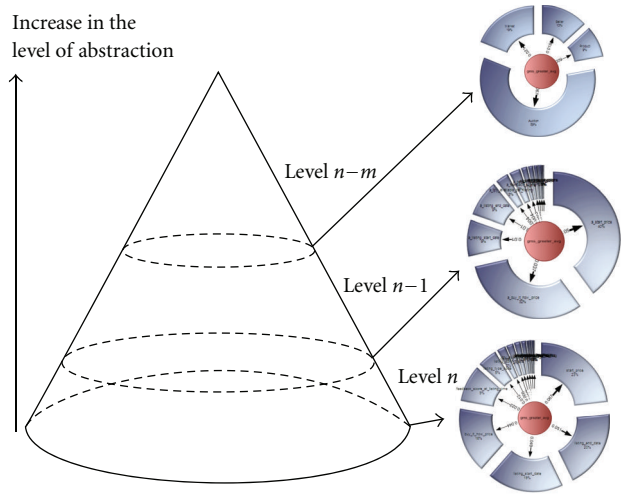

Figure 8: Concept hierarchy and multilevel h-data cone.

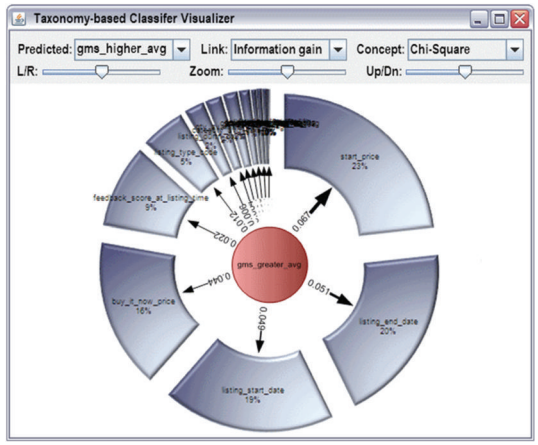

Figure 9: Visualization of attribute-to-class information at level n (most bottom layer).

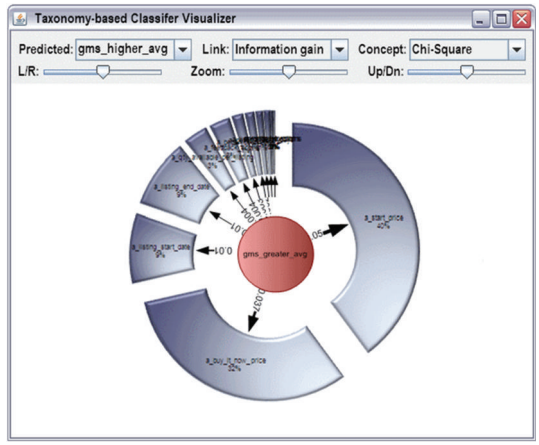

Figure 10: Visualization of attribute-to-class information at abstract level 3 (middle layer).

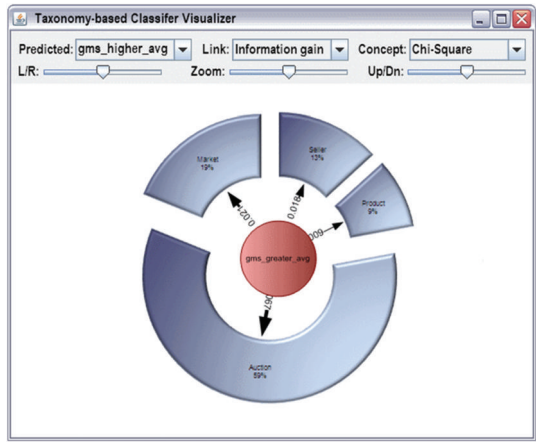

Figure 11: Visualization of attribute-to-class information at abstract level (high layer).

(3) Attribute information: (circle diameter)

correlations to the target class,
correlations to the other attributes
worthiness of attributes (by Chi-Sq. algorithm).

Some snapshots of the visualization are shown in Figures 9, 10, and 11. They display the information associated with the attributes that are increasingly abstracted from Figures 9 to 11. Biometrics authenticators therefore have the flexibility of utilizing the interrelation information of attribute-to-attribute and attribute-to-class at different abstract views for formulating questions.

One interesting observation is that the visualized charts indicate that the blood glucose concentration has the most influential factor in predicting the abnormal conditions. By this information from the h-data, the authenticator may question the test subject about his average blood glucose concentration while his abnormal conditions are already known. However, this may be a very well-known fact because the abnormal conditions are derived from the BG measurements. So the authenticator may want to turn off the attribute group BG and continue to search for the next greatest predictive strength of other attribute groups for formulating more challenging questions. The other observation is that in Figure 10 when the attributes are abstracted into major concept, at a glance we can see that neuroglycopenic symptoms relate to concepts of the following order: insulin, light diet, and heavy exercise. The concept is an abstract form that embraces all the life-style patterns related to the blood glucose concentration. So questions about the test subjects lifestyles in terms of diets and amount of daily exercise may be asked. The last resort for authentication is of course a small blood test for collecting his actual insulin and glucose level. But with the h-data, we have the flexibility of deriving authentication questions from simple (general) to complex by descending along the hierarchy.

The model we adopted here will work best when there are many attributes from which meaningful concepts can be abstracted. Also the AVT is good to have many distinctive

levels, thus many levels of resolutions can be generated for use in question searching upon authentication. Some common levels of resolutions that we encountered from attributes of datasets in data analytics include:

continent ⟶ country ⟶ province ⟶ city ⟶ street
year ⟶ sason/quarter ⟶ month ⟶ week ⟶ day
population ⟶ clan ⟶ body ⟶ organ ⟶ cell.

6. Conclusion

Biometric authentication in the past has taken many forms of unique bodily features. In this paper a novel concept of biometric authentication by exploiting a user's medical history is proposed. Similar concepts have been raised recently by using information about the user's unique online activities and email logs. However, medical history is relatively stronger than activity events because each medical event is supposedly verified by medical professionals—the records can be traced, the medical history can hardly forged and instant testing can be made available (when necessary) by a body examination on the spot. The application of medical history in user authentication is suggested to assume a question-based form; few short questions must be answered by the testing subject upon authentication. Direct questioning is believed to be inappropriate because users may be reluctant to confess his medical conditions especially in front of a human validator, and security of the medical history may be comprised if they are used explicitly in the authentication process. Hence, in this paper we stress on a need that authentication should take on an implicit form such that users will no longer have to be confronted with his medical conditions. Instead general questions such as his lifestyle and dietary habits would be asked whose answers will be then inferred to the priori answer (the illness and its extents, etc.) for authentication matching. To facilitate such implicit questioning, a new type of data representation, namely h-Data is introduced. h-Data has a hierarchy of resolution for defining the information about the medical condition. A biometric security card can store a number of h-Data, corresponding to each of the user's medical illness if he ever suffered from multiple major illnesses. Essentially each layer of h-Data is a relation-map that maps the attributes while specifying their relations and their strengths to the target class. With h-Data the authenticator can have the flexibility of gliding along the hierarchy in search of questions ranging from general to specific. Because the medical conditions are already known, inferring from the answers to those general questions can lead to a hypothetical answer (medical condition) that could be used to test if it matches with the actual one stored. This paper contributes to the original idea which is believed to be the pioneer in using medical history for user authentication. What follows will be extensive research from the authors and hopefully from the scientific community to further perfect this technological innovation. Many future areas revolving this concept exist, such as applying natural language in deriving authentication questions, security and usability evaluation, and accuracy testing of the said technology, hardware and software system design, messaging protocols, and so forth, just to name a few.

References

[1] A. Nosseir and S. Terzis, "A study in authentication via electronic personal history questions," in *Proceedings of the 12th International Conference on Enterprise Information Systems (ICEIS '10)*, vol. 5, pp. 63–70, June 2010.

[2] M. Nishigaki and M. Koike, "A user authentication based on personal history- a user authentication system using e-mail history," *The Journal on Systemics, Cybernetics and Informatics*, vol. 5, no. 2, pp. 18–23, 2007.

[3] "SmartMetric Announces Its Fingerprint Biometric Card Can Now Be Used to Hold Personal Medical Records Without Security Compromise," August 2010, http://www.smartmetric.com/.

[4] "New biometric security card in force at Canada's largest airports," Wings Magazine, Annex Business Media , 2011, http://www.wingsmagazine.com/.

[5] S. A. Samad, D. A. Ramli, and A. Hussain, "A multi-sample single-source model using spectrographic features for biometric authentication," in *Proceedings of the 6th International Conference on Information, Communications and Signal Processing (ICICS '07)*, pp. 1–5, December 2007.

[6] M. Demel and G. Ecker, "New challenges for feature selection: on the relationship between feature selection and classification accuracy," in *Proceedings of the JMLR Workshop*, pp. 90–105, 2008.

[7] J. Zhang and V. Honavar, "Learning decision tree classifiers from attribute value taxonomies and partially specified data," in *Proceedings of the 20th International Conference on Machine Learning*, pp. 880–887, August 2003.

[8] H. Jo, Y. C. Na, B. Oh, J. Yang, and V. Honavar, "Attribute value taxonomy generation through matrix based adaptive genetic algorithm," in *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '08)*, pp. 393–400, Dayton, Ohio, USA, November 2008.

[9] M. Guo, J. B. Yang, K. S. Chin, H. W. Wang, and X. B. Liu, "The evidential reasoning approach for multi-attribute decision analysis under interval uncertainty," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 3, pp. 683–697, 2009.

[10] Z. Lin, M. Hewett, and R. B. Altman, "Using binning to maintain confidentiality of medical data," in *Proceedings of the American Medical Informatics Association Annual Symposium (AMIA '02)*, pp. 454–458, 2002.

[11] S. Dawara, *Grouping related attributes*, M.S. thesis, The Rochester Institute of Technology, 2004.

[12] G. Das, H. Mannila, and P. Ronkainen, "Context based similarity measures for categorical databases," in *Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD '00)*, pp. 201–210, September 2000.