A MULTI-LEVEL BIOMEDICAL CLASSIFICATION MODEL BY USING AGGREGATION AND ABSTRACTION TECHNIQUES

Simon Fong
Department of Computer and Info. Science
University of Macau
Macau
ccfong@umac.mo

Andy Ip
Faculty of Science and Technology
University of Macau
Macau
henry.yh@umac.mo

Sabah Mohammed
Department of Computer Science
Lakehead University
Thunder Bay, Canada
sabah.mohammed@lakeheadu.ca

ABSTRACT
Data mining on biomedical data usually faces challenges of preserving privacy and finding associations among the attributes. Comprised of various and meticulous clinical measurements, the data to be data-mined often carry many attributes. When all these attributes are used in constructing a classification model, it may lead to a well-known problem in data mining called over-fitting which results in poor prediction accuracy. At the same time, the high resolution (or details) of the attributes may compromise the privacy of the patients’ identities. In this paper a multi-level classification model is proposed to analyse biomedical data with the attributes flexibly abstracted and aggregated at will of the user. The novel method contributes to biomedical research community in threefold: (1) increasing the prediction accuracy; (2) subsiding the privacy issue; and (3) enabling the relations between the attributes of the data to be further analysed by biomedical experts. The prototype of the model is tested via several experiments with some classical biomedical data obtained from UCI. A visualization tool is also programmed that shows both the significances of the attributes and their predictive powers. The experimental results indicate that by applying appropriate aggregation and abstraction techniques, decision trees can make to be more compact and more accurate.

KEY WORDS
Biomedical data classification, Decision tree, Multi-level data visualization.

1. Introduction

Recent advances in bio-technologies such as those in gene-functions studies, pharmaceutical studies, medical imaging, clinical applications and electronic medical records etc are generating a tremendous amount of data, often stored in heterogeneous formats and in distributed bio-medical databases. This accumulated volume of biomedical data that waits for in-depth analysis usually carries a wide variety and large amount of attributes. It was already pointed out in [1] that appropriate data mining techniques are needed for data cleaning and data pre-processing, as an important task for systematic and coordinated analysis of bio-medical databases.

The data attributes, from an analytic perspective, will evolve to become more complex than ever, due to the advances but uncontrolled data collection and generation methods, that will lead to the ill effect in data mining by the phenomenon called Curse of Dimensionality [2]. It refers to the exponential growth of the high-dimensional space as a function of dimensionality. That will have ill impacts on induction-based decision tree algorithms (such as C4.5 and its variants) as the tree nodes and leaves would escalate exponentially in heap memories of a computer by the growth of dimensions. The accuracy and comprehensibility of the trees will diminish as a consequence, which is known as over-fitting. One extreme example is a drug discovery dataset called Dorothea, provided by DuPont Pharmaceuticals and it is made available for download at UCI Data Repository (http://archive.ics.uci.edu/ml/index.html). Dorothea has 100,000 attributes taken from tests of many small molecules for their ability to bind to the target site, in order to determine what separates the active (binding) compounds from the inactive (non-binding) ones.

Some typical preprocessing tasks in addition to tree pruning, have been widely applied. They include: Data cleaning - Removal or substitution of missing values, noisy data, and outliers, Data integration - Integration of multiple databases, data cubes, or data sources, Data transformation - Normalization and aggregation of attributes and their values, Data reduction - Feature selection and sampling, that down scale the data in columns (dimension) and in rows (volumes) respectively, and Data discretization - Reduction of lengthy numerical data into discrete categorical values. Although these techniques have their own merits in certain aspects and at times they are used collectively in pre-processing, they are rather based on quantitative characteristics of the attribute values than based on the meanings of the attributes. Hence attributes are combined, transformed or omitted without referencing to their ontological meanings. For example, 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 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 multi-resolutions of taxonomy must be maintained.

Attribute value taxonomies (AVT) that was proposed by Honavar [3], 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 [3, 4] is formulating a new breed of learning classifiers, namely AVT-Decision Tree for deriving rules directly from AVT’s that are constructed from the data.

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.

![Sample AVT for date attribute of a dataset](image)

Figure 1. Sample AVT for date attribute of a dataset

This approach is especially useful when we deal with data whose attributes have complex contextual resolutions. Taking clinical data records as an example, 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, e.g. 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 1, but they belong to different concept groups, placed in different levels. Researchers may be interested in knowing the inter-relations among the attributes at different abstract level, and in relation to the predictive power in a decision tree.

On the other hand, by generalizing and grouping attributes and their values to specific concept levels, the anonymity of the data can be enhanced. Biomedical 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 for rare cases. 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.

In this paper, we devise a biomedical classification 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 visualized in terms of their predictive powers to the target class, and other inter-attributes 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 biomedical ontological databases. We used a collection of biomedical datasets as a case study, for evaluating the feasibility of the model.

2. Related Works

Several underlying techniques such as attribution selection and grouping, data aggregation, and Attribute Value Taxonomies have been exploited extensively in the literature. These underlying techniques that serve as core components in our model are reviewed in this section.

Honavar et al first proposed AVT in 2003 [3], which is a structured taxonomy defined over values of an attribute. AVT-guided decision tree learning algorithms were proposed in [3, 5] that exploit a taxonomy of propositionalized attributes as prior knowledge to generate compact trees. Also they utilized genetic-algorithm to automatically construct taxonomy from data [4], given that some taxonomies are unavailable in some application domains. AVT would be used as a useful reference guide for data preprocessing.

Feature selection and dimensionality reduction concern about reducing the attribute space of a feature set; the former deals with choosing a subset of original attributes, and the latter attempts to produce linear combinations of the original attribute set. The relationship between attribute space reduction technique such as Principle Component Analysis (PCA) and the resulting classification accuracy was investigated in [6]. It was concluded that the classification accuracy by PCA differed across different types of data, and the variance in PCA was not a vital factor in the classification performance. Another popular approach is grouping attributes by measuring their distances apart in the hyper-plane. Dawara [7] argues that traditional clustering techniques fail to adequately group attributes into correct taxonomical groups because most of the clustering
algorithms are based on measuring the mathematical distances based on the values of the attributes, instead of the ontology of the attributes.

The past works purport that despite the numeric characteristics in the attributes there may exist other factors to the classification accuracy. Measuring the statistical aspects of the attributes alone may not be sufficient to group the attributes into a hierarchy of concepts that can be manipulated in different resolutions.

Yang and Sanver [8] advocate the importance of having a common representation of aggregated data for analyzing a large dataset that is not only large in volume but contains many attributes. Such a data representation is needed to be in multiple resolutions. This makes data analytic tools more scalable to large data sets. E.g. For location, we can have: Continent ➔ Country ➔ State ➔ City ➔ Street; for date, we may have Year ➔ Month ➔ Week ➔ Day. The work [8] however is geared towards database technology, though the authors suggested that Multi-resolution data aggregation brings unique opportunities to data mining algorithms.

Another important related work is [9] by Legrand and Nicoloyannis who proposed a feature selection method based on preference aggregation. The method uses preferences aggregation to determine an ordered list of features subsets. The criteria in the preferences aggregation can be flexibly defined by users for finding an optimal features subset by filter-and-wrapper approaches. The logics of the algorithm can be extended to work on AVT with multi-resolution levels.

We take a step further to apply the concept-hierarchy in data preparation and then supplying the transformed data to build multiple decision trees (DTs). After the DTs are built, we retain the performance data and use these data for visualizing the relations among attributes and attributes-to-class.

Specifically for biomedical data management, Altman et al [11] proposed a binning algorithm to ensure that no unique records are available to the users. The data are generalized upwards in hierarchies until the values of records are shared by a user-specified number of records, called the bin size. In this way, the identifying information is removed from medical and genomic data. A parameter is defined to control the trade-off between privacy and data integrity. The bigger the bin size is, the more difficult for any record to be traced to a single person, yielding more privacy. Otherwise, smaller bin sizes result in more specific and detailed data. All symbolic and numeric data are represented hierarchically, and they are binned into generalized tree nodes. The different levels of precision relate to the trade-off between privacy and data resolution. We adopted this concept of hierarchy tree and data generalization basically, and infused it to the classification model coupled with visualization and feature selection pertaining to biomedical data with high dimensionality.

3. Multi-level Classification Model

The framework of the multi-level classification model is shown in Figure 2. The central components in the framework are the aggregation and abstraction processes after the pre-processing mechanism that receives a set of raw data as input. The raw data are transformed into several datasets prior to tree building process. The original dataset has all the attributes. The other important input file is a predefined concept hierarchy represented in AVT format by domain expert. The input of the concept hierarchy also specifies the number of levels and what the subgroups in each level are.

The transformed datasets are the data that have been aggregated and abstracted 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 $M = M_1 \geq M_2 \geq M_n$ where $M$ is the original dimension of the initial dataset, $M_1$ is the new dimension of the transformed dataset $L_1$ at level 1, $M_n$ is the root of the AVT which also is the highest level, $L_1$ is the dataset that has the $M$ number of original attributes.

With the transformed datasets $L_1$ to $L_n$, traditional tree building process e.g. C4.5 induces the corresponding decision trees, DT to $DT_n$ as outputs. Because of the reduced dimensionality the sizes of the trees follow this pattern: $C(DT_1) \leq C(DT_2) \ldots \leq C(DT_n)$ where $C(DT)$ is the size of the DT in terms of the sum of nodes and leaves. Once the $DT_{1,2,n}$ 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 pre-processing 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 viewers 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. Our model offers the flexibility for users to choose the level of resolution at which the information about the attributes (attributes-to-attributes or attributes-to-target-class) to be viewed. For an example of a clinical application, whether and how much a seasonal climate or some general patterns of lifestyles would affect the effectiveness of a medical prescription, may make more sense and better interpretable than reading the detailed measurements of a group of patients or very specific information on the individual attributes.
As shown in Figure 3, the DTs information can be visualized at different levels of resolution corresponding to the cone-shaped concept hierarchy. At the top the attributes are most abstracted, and the bottom is the full set of original attributes. We can see that the multi-level information visualizer works on different sets of DTs information which are derived from transformed datasets of a particular level. During the abstraction process, the input AVT was used to guide the abstraction accordingly to certain levels and certain groupings of attributes. The user at ease can choose which sets of DTs information of different resolution that he/she wishes to visualize about.

There are basically three components that construct a visualization image by our design. First is the center dot that is the predicted class or the target attribute a decision tree predicts about. Second is the link information in the shape of an arrow with varying thickness. For viewing the attribute-to-class relations, each arrow uses its thickness to represent the predictive power of that particular attribute (or a group of attributes, depending on the abstraction level) pertaining to the predicted class (or sometimes called target class in data mining). The thicker the arrow is, greater the predictive power the attribute possesses. The third component is item size, which looks like a portion of a pie in Figure 4. The item size is scaled proportionally to the correlation value with respect to the target class (which is the center dot). So the more correlated an attribute is, relatively to the target class in percentage, the larger the item. The types of information, such as predictive power, information gain or correlation value can be interchanged between the Link Strength and Item Size. Users are able to select the choices of information representations from the user-interface.

Likewise, users can replace the center dot from a target class to any attribute available from the list. By doing so, the visualization is switched from attribute-to-class to attribute-to-attribute. This way, users can view at will the relations between any pair of attributes.

In order to facilitate such design, the abstraction and aggregation methods would have to generate all the necessary combinations of transformed datasets. All the corresponding DTs and the associated statistic results must be computed and stored up prior to the visualization. Readers who are interested in the technical details are referred to [13].
4. Aggregation Methods

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.

Two examples occurred in our case study of analyzing UCI biomedical data. 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 the values across a number of the measurement attributes.

Figure 5. Examples of aggregation on binary variables in UCI biomedical datasets

The other example is on aggregating a set of conditional attributes that have binary values (true or false) into a single attribute. In the UCI biomedical 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 for multi-attribute decision analysis [11] is applied to estimate the ranks. Some other examples are shown in Figure 5.

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, we adopted a combined approach of binning and histogram analysis. The data are categorized by quartile analysis over a normal distribution of frequency. The quartiles (25% each) is used to grade the new ordinal variables as Low ≤ Q1, Medium_low ≤ Q2 and Q1, Medium_high ≤ 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.

5. Abstraction Methods

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 to a concept found in the next higher level. The process repeats until all the concepts are done, level by level in the AVT. The data format for representing an AVT would take the following form, similar to that in [9].

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

\[
\text{avt}(\text{number of concept, <concept names>})_{\text{level number}} = \text{avt}(1, \langle\text{all diabetes records}\rangle)_1, (4, \langle\text{insulin, glucose, exercises, diet}\rangle)_2, \ldots (M, \langle\ldots\rangle)_n).
\]

where

\( M \) is the number of attributes, \( a_i \) in level \( i \)
\( L_i \) is the working dataset in level \( i \)

Dataset \( L \) can be viewed as a two-dimensional matrix such that \( L_i = D_i (M, R_i), i = 1 \ldots n \). Let \( m_{var} = M \) and \( r_{var} = R_i \), in level \( i \). A dataset in \( D \) has \( m \) attributes i.e. \( D = (a_{1,i}, a_{2,i}, \ldots a_{m,i}) \) with \( R_i \) instances in level \( i \) of \( avt \).
In [13], the function \( \text{Abs}(D_i, L_i) \) is to partition attributes \( a_{i_1}, \ldots, a_{i_m} \) 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 describe a common concept. The clusters themselves may be relatively different from each other. Therefore less clusters or concepts would be found in a higher level up – 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 \( a_i \) is not available, i.e., we base solely on the information of the attributes and their values to form clusters. When the \( a_i \) 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 [7], is to measure the distances of the concepts and to determine how the concepts should be grouped by the attributes, should \( a_i \) 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 biomedical record and assume somehow we have some missing information or uncertainty in a level of concepts in the \( a_i \), 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.

### 6. Experiment and Results

In order to verify the multi-level biomedical classifier model presented above, a number of data set were used in experiments to test out the outcomes. The biomedical 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 auto-immune attack or 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 is diagnosed and is receiving regularly scheduled exogenous (externally administered) insulin, the principal metabolic effect of concern is the potential for hyper-glycemia (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 patients’ historical records of their relevant diabetics conditions, a classifier should be able to forecast the blood glucose level based on the values of the other measurement attributes. In other words, by means of the training data a suitable model for prediction, whether the predicted glucose level deviates from the safety limit, should be developed. We can see that the prediction problem is somewhat complex because many attributes may contribute to the prediction target up to certain extent. And each of the inter-relations 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 prediction problem, a multi-level classification model is to be built. Firstly, we attempt to model an AVT 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 Figure 6 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 visualizing the relations of the clustered attributes to the predicted class.

![Figure 6. Relationship diagram of abstract groups](image-url)
The prediction goal is defined by two objectives, 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:

Abnormal Blood Glucose (BG) Conditions:
- Pre-meal BG falls out of ranges 80-120 mg/dl
- Post-meal BG falls out of ranges 80-140 mg/dl
- 90% of all BG measurements > 200 mg/dl and that the average BG is over 150 mg/dl

Hypoglycemic (low BG) symptoms:
- Adrenergic symptoms, BG between 40-80 mg/dl
- Neuroglycopenic symptoms, BG below 40 mg/dl

The end result is a collection of decision trees, with each specially prepared for the abstract concept views of a level in the AVT. What may be even more interesting to the analysts is the visualization of the attributes in different abstract views. The visualization prototypes are programmed in Prefuse\(^1\) which is an open-source interactive information visualization toolkit, and Java 2D graphics library. Through the select menu in the graphical user interface, analysts have the options of choosing to see the combinations of the three domains of information:

1. Predicted class: Abnormal BG, Pre-meal (center circle)
   - Abnormal BG, Post-meal
   - Abnormal BG, General
   - Hypoglycemic, High BG
   - Hypoglycemic, Low BG

2. Link strength: Predictive power to the target (line thickness)
   - Rank of relevance to the target
   - Info. gain w.r.t the target

3. Attribute information: Correlations to the target class (item size)
   - Worthiness of attributes (Chi-Sq.)

Prototypes of the visualization are shown in Figures 9 and 10. They display the information associated with the attributes that are increasingly abstracted. Biomedical analysts have the flexibility of viewing the inter-relation information of attribute-to-attribute and attribute-to-class at different abstract views.

One interesting observation is that the visualized charts indicate the blood glucose concentration has the most influential factor in predicting the abnormal conditions. This is a well-known fact because the abnormal conditions are derived from the BG measurements. In this case, one would want to turn off the attribute group BG, and continue to investigate the predictive strengths of other attribute groups. The other observation is that in Figure 9 when the attributes are abstracted into major concepts, 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 lifestyle patterns related to the blood glucose concentration.

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 in the visualization. Some common levels of resolutions that we encountered from attributes of datasets in data mining include:

- Continent \(\rightarrow\) Country \(\rightarrow\) Province \(\rightarrow\) City \(\rightarrow\) Street
- Year \(\rightarrow\) Season/Quarter \(\rightarrow\) Month \(\rightarrow\) Week \(\rightarrow\) Day
- Race \(\rightarrow\) Group \(\rightarrow\) Entity \(\rightarrow\) Body \(\rightarrow\) Organ \(\rightarrow\) Cell

\(^1\) http://prefuse.org/
Table 1  
Performance Comparison Chart of DTs by Diabetics Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>All original attributes are used</th>
<th>Attributes are aggregated</th>
<th>Attributes are aggregated and</th>
<th>E-Measure</th>
<th>No. of leaves</th>
<th>Size of tree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>73.83%</td>
<td>73.36%</td>
<td>74.67%</td>
<td>0.735</td>
<td>20</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.735</td>
<td>40</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2  
Accuracy of Decision Trees by Various Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Prediction Accuracy</th>
<th>Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All attributes</td>
<td>Aggregation and Abstraction</td>
</tr>
<tr>
<td>Breast-cancer-wisconsin</td>
<td>94.42%</td>
<td>95.42%</td>
</tr>
<tr>
<td>Dermatology</td>
<td>79.24%</td>
<td>85.52%</td>
</tr>
<tr>
<td>Pima-indians-diabetes</td>
<td>84.11%</td>
<td>85.18%</td>
</tr>
<tr>
<td>Breast-cancer_UF</td>
<td>95.66%</td>
<td>95.65%</td>
</tr>
<tr>
<td>Hepatitis</td>
<td>62.56%</td>
<td>85.45%</td>
</tr>
</tbody>
</table>

Figure 9. Visualization of attribute-to-class information before aggregation and abstraction  
(with Link strength = Predictive power to the target, and Item size = Correlation to the target class)

Figure 10. Visualization of attribute-to-class information after aggregation and abstraction  
(with Link strength = Predictive power to the target, and Item size = Correlation to the target class)
The performance information of the different decision trees are collected, and they are put vis-à-vis into comparison as shown in the Table 1. We can see a general trend that when the original attributes are transformed by aggregation and abstract accuracy improves and the decision tree becomes more compact in size. It indicates that the effect of over-fitting was being subsided by downsizing the trees. At best when the tree is built from abstracted concept, a high accuracy of over 0.8 F-Measure is achieved. Table 2 shows a general trend that the prediction accuracy of the multi-level decision trees built by other biomedical datasets from UCI, has improved when compared to the one with all the original attributes.

7. Conclusion

Biomedical data are inherently complex and comprised of many attributes that may not be easy for constructing a compact decision tree model for classification and prediction. Various techniques that include feature selection and data transformation have been studied in the past, with the aim of producing a compact and efficient decision tree. They all have their respective strengths, but in general they commonly lack of preserving the meanings of the attributes. The concept of Attribute Value Taxonomies (AVT) that is a value set of attributes specified at different levels of precision in a tree-structure was originally proposed by Honavar in year 2003. AVT has the advantages of naturally and easily understanding the attributes in a hierarchy of resolutions. In this paper, we extended the concept of AVT into the domain of biomedical data classification by building decision trees based on attributes that are abstracted in different levels. The output of the process is a series of decision trees with each specifically built pertaining to an abstract level of concept, which we called it, a multi-level classification model. A collection of biomedical dataset from UCI was used as experiments for evaluating the performance in terms of accuracy. Diabetics dataset was used as a case study for illustrating the merits of the model. Other datasets were run and the prediction accuracy of the respective decision trees generally shows improvement. The results are more compact decision trees, and the inter-relation information between the attributes and class can be visualized at ease. The accuracy of the decision tree after the attributed have been aggregated and abstracted has a significant improvement. The resultant trees can be mapped across to AVT for easy interpretation by human.

References

[1] J. Han, "How Can Data Mining Help Bio-Data Analysis?", BIKDD02: Workshop on Data Mining in Bioinformatics (with SIGKDD02 Conference), Extended Abstract.


