A Decision Support Library Using Multiscale Modeling And Disease State Index Method For Prognosis

Ramya R B
PG Scholar
Dept. of Computer Science And Engineering,
Anna University,Regional Center
Coimbatore

D. Palanikkumar
Assistant Professor,
Dept. of Computer Science And Engineering
Anna University,Regional Center
Coimbatore

Abstract

Information overload and lack of adequate time to review all the documents is a grand challenge faced by physicians while making a diagnostic decision.This facilitates the need for the modeling of a clinical decision support system where the physician does not have to review the piles of documents.Diseases like cancer require the identification and characterisation of genetic and molecular properties of cells and their dynamic interactions.Multiscale relates to patient data measured at multiple scales like structural, atomic, genetic, molecular, neuro-pshycological etc. The objective is to develop a composite representation by integrating all these models in such a way that the relevant information about the patient's disease state will be highlighted and having a glance the clinician can interpret and make a diagnostic decision easily.The objective of this project is to design a clinical decision support system (CDSS) that supports heterogeneous clinical decision problems using multiscale modeling.Meeting this objective required a novel design to create a data –agnostic CDSS for point of care support.The proposed solution is evaluated in a proof of concept implementation.

KEY WORDS

Decison Support System, Multiscale,Supervised Learning, Support Vector Machine

1.Introduction

We know in the field of medical research and clinical practice, clinicians have to deal with huge amount of data. Beginning from the questionnaire that occurs when we first consult a doctor to the laboratory results and results of sophisticated imaging techniques, the clinicians have to consider voluminous data before taking a diagnostic decision. This is a time consuming and frustrating task. In addition to printed materials, physicians have to review electronic journals, websites, RSS feeds, streaming videos, and blogs.Having so much information that it becomes confusing and overwhelming can effect patient outcomes.

"Information overload" is a term used to describe the difficulties one can have when there is so much information that it is impossible to review it all before making a diagnostic decision. In addition clinicians often report they do not have enough time to pursue answers or use new innovations. In order to rectify the issues that arise due to information overload and medical errors computer-based clinical decision support systems have been developed.The idea of using CDSSs is to provide improved health care at the point of need at reduced costs.

Clinical Decision Support Systems (CDSS) have been developed based on different methodologies. Any computer software or program that simplifies or aids in making a clinical decision can be categorised as a CDSS. Simulating disease behavior across multiple biological scales in space and time, i.e., multiscale modeling, is identified as a powerful tool for generating accurate predictions. These models consider biophysical, biochemical, and biomechanical factors .

In physiology, there are several significant motivations for multi-scale modeling. First, the need to model at multiple levels – spatially and temporally – becomes evident by simply considering physiology: organisms consist of organs, which are made of different tissues; these tissues are in turn made up of sets of specialized cells which interact with each other. Complex biochemical processes regulate cell metabolism and short and long term behaviour within the cells. All these activities take place on

the underlying protein kinetics and gene expressions.There prevails a strong interconnection between all the levels and the normal functioning at a given level depends not only on the underlying synergy of sublevels but also on higher levels, such as feedback via neural and hormonal mechanisms. Multi-scale patient-specific modeling is a most promising, innovative research area in computational biology and possibly the future of medicine.

2. Related work

A novel generic clinical decision support system is developed, which models a patient's disease state statistically from heterogeneous multi scale data. Its objective is to aid in diagnostic work by analyzing all available patient records and highlighting the relevant information to the clinician[1]. The system is assessed by applying it to several medical datasets and demonstrated by implementing a novel clinical decision support tool for prognosis of Alzheimer's disease but a significant disadvantage is that it does not support complex values and value lists and is currently implemented for a single disease.

Cancer is a class of diseases characterized by malignant growth ie., abnormal and uncontrolled cell division and tissue invasion and it may spread to other parts of the body through the lymphatic system or the blood stream.So it is a problem scenario where multiscale modelling is necessitated.The atomic scale serves the purpose of studying the structure and dynamic properties of proteins, peptides, and lipids, as well as their dependency on the features of the environment or on ligand binding.Models at the molecular scale do not represent individual proteins, but they represent an average of the properties of a population of proteins. Cell signalling mechanisms ie., the natural regulators of biological systems, are usually examined at this scale.The microscopic scale is also referred to as the tissue or multicellular scale, it also includes the cellular scale which encompasses single-cell behaviors and properties. Individual cells are contained in a selectively permeable cell membrane. In the context of cancer, models at this scale describes the transformation of normal cells into malignant ones, associated alterations of cell-cell and cell-matrix interactions and the heterogeneous tumor environment.Models at the macroscopic scale focus on the dynamics of the gross tumor behavior, including shape, extent of invasion,morphology etc. under different environmental conditions[2].

Support Vector Machines, one of the new techniques for pattern classification, have been widely used in many application areas.When using SVM, two problems are confronted: how to choose the optimal input feature subset for SVM, and how to set the best kernel parameters.These two problems are crucial, because the feature subset choice influences the appropriate kernel parameters and vice versa[3].Therefore, obtaining the optimal feature subset and SVM parameters must occur simultaneously without degrading the SVM classification accuracy. A genetic algorithm approach for feature selection and parameters optimization to solve this kind of problem is presented.The basic idea of the GA-SVM method is to remove the features which are less fit. The features that have high fitness value and high classification accuracy are retained for the evolution.This is achieved by an iterative algorithm. Our proposed GA-based approach significantly improves the classification accuracy and has fewer input features for support vector machines.

3. Proposed System

A data-agnostic statistical disease modeling method has been developed which combines heterogeneous multiscale data to compute a value in the interval [0,1].This value indicates a patient's disease state, i.e., the location or rank based on data, in relation to previously known control (healthy) and positive (disease) populations.It can be viewed as a supervised classifier, where patient data are compared to previously diagnosed data.

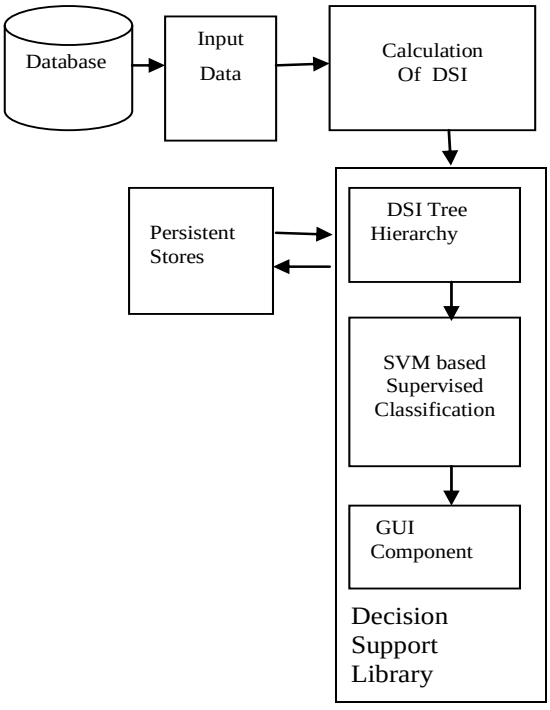

Figure 1. Architecture of the Proposed System

The workflow is as follows:

1.Compute the Disease State Index,Fitness and Relevance functons.
2.In the case of complex values and value lists modify the functions appropriately.
2.Perform recursive evaluation of these functions.
3.Represent the DSI tree hierarchy.
4.Transform the data to the format of an SVM package and conduct simple scaling on the data.Consider the RBF kernel and use cross-validation to find the best parameters and then use the best parameters to train the whole training set and test.
4.Use Genetic Algorithm to get the best fit values and select a certain number of cases with high fitness value then use SVM to predict the target value of test data.
5.Calculate the accuracy of classification in case of both SVM and GA-SVM.
5.Render the results using the GUI component.

3.1 DSI Evaluation

Given the heterogeneous patient data from a single test at a single time point, e.g., an individual neuropsychological test or laboratory analysis results of a blood sample, as $x_1,x_2..$, disease state index (DSI)function is defined as a weighted mean .

$DSI(x_1, x_2, \ldots . x_n) := (\sum_{i=1}^{n} Rel(i) * Fit(xi)) / (\sum_{i=1}^{n} Rel(i))$

where Rel(i) is a relevance function providing the weighting between [0,1] for variable i and Fit (x_i) is a fitness function providing a nonlinear transformation of value x_i into fitness space [0,1].A fitness function computes the location, i.e., rank, of an individual variable x_i relative to values of the same variable in two different populations, denoted as controls C_i and positives P_i.Consider a scalar variable where the progression of a disease tends to increase its value. For these, fitness is defined as a monotonically increasing function

$Fit(x_i) := LP(x_i))/(LP(x_i)) + RC(x_i))$

where LP (x_i) is the left integral of probability density function (PDF) for positive class values P_i and RC (x_i) is the right integral of PDF for control class values C_i . Derivation of the fitness function can be conducted in an analogous manner for ordinal variables. For a categorical variable $x_i \in \{\Omega_1,.. ,\Omega_n\}$,the conditional probability of the subject belonging to the positive population in the case of observing $\Omega = x_i$ is used as the fitness function.In case complex values and value lists are the input data these functions have to be modified appropriately,probably using an absolute function wherever x_i is used.

The weighting factors of DSI, i.e., relevancies of variables, are determined by the variables' ability to correctly classify between the known classes C_i and P_i , and are independent of the patient data. Relevance is defined for scalar and ordinal values that increase with disease progression as

$Rel(i) := \max\{0, LC(x_i^*) + RP(x_i^*) - 1\}$

where LC (x_i^*) is the left integral of PDF for control values Ci and RP (x_i^*) is the right integral of PDF for positive values Pi at the decision threshold x_i.For categorical variables, relevance is the classification accuracy of training cases given the category of the independent variable.

To combine data from multiple tests and/or multiple scales, DSI values obtained are recursively inserted back into as new variables, thereby using several levels of recursion for granularity.Recursive evaluation provides fitness, relevance, and DSI values for a tree of data.The leaves and branches represent multiple scales but converge to a common root which the whole system. This tree of data can be rendered for quick visual interpretation of multiscale data, using colors and shapes to quickly make out patient state and the relevance of all tests and variables. The nodes can also be ordered according to their relevance to show the most significant features at the top .Larger node sizes indicate higher relevance (i.e., better discrimination of Training classes), with irrelevant features omitted. Shades of red signals the similarity of the patient data to the disease population , shades of blue similarity to healthy.

In summary, DSI uses available multiscale data to model the state of having a disease. It does so first with the individual measurement values, then transforms the values nonlinearly to a common classification space and combines them within that space to obtain aggregate results. The recursive computation produces classification results at multiple levels of abstraction.This can be visualized using a tree hierarchy where the visualization clearly discloses how patient data contributes to the disease state, facilitating rapid interpretation of the information.

3.2 Decision Support Library

A software library implementing the DSI computational method and supporting features using the Java language is developed .The library is context independent, and hence is applicable to several domains.The library supports accessing multiple data repositories with a layered approach since the DSI can use any available multiscale data. Data access implementations, called persistence stores are free to connect with data sources in any way that is through an object relational mapping (ORM) service, web services, or simply reading a flat text file. An interface defines how the

persistence stores can transfer data to and from the library.

A data definition layer consists descriptions of entries (e.g., types of tests done to a patient) and feature values (types of individual data points) within those entries.The organization of the DSI tree hierarchy is also described within this layer in addition to all features existing at the leaf nodes.The actual data that are analyzed are contained within another layer, within which all the subjects, entries, and feature values are represented by matching object instances.In order to perform DSI computations control and positive classes are to be constructed in a generic manner, using entities from one or more persistent stores that provide training data.

On having the training data, data from the patient we are studying, and with the definition layer describing the feature hierarchy, the library has all the necessary information for evaluating the DSI.Training data obtained is organized in the form of a tree hierarchy where the leaves contain actual measurement values for the training set. Fitness and relevance are evaluated at the leaf level, DSI and relevance values in internal nodes are computed recursively, and, finally a total DSI value for the whole dataset at the root of the DSI tree is obtained.

The supervised machine learning approach has been widely applied to bioinformatics and gained a lot of success in this research area. With this learning approach based on SVM researchers initially develop a large training set.Active learning is performed with support vector machine and the algorithm is applied to gene expression profiles of prostate cancer,breast cancer and lung cancer samples. Support Vector Machine algorithm is used to make the classification.The point of SVM is to produce a model (based on the training data) which predicts the target value of the test data given only the test data attributes.With SVM each data instance is to be represented as a vector of real numbers.Also if categorical attributes are involved, some preprocessing is necessary ie., they have to be converted to numeric data.Scaling prior to application of SVM is important and the main advantage is to avoid attributes in greater numeric ranges dominating those in the smaller ranges. Another advantage is to avoid numerical difficulties during the calculation.For achieving a relatively high prediction accuracy we have to use the same method to scale both training and testing data.

Though there are four common kernels, in general, the RBF kernel is a reasonable first choice.The RBF kernel maps samples non-linearily into a higher dimensional space.In contrast to,the linear kernel, it can handle the instance when the relation between class labels and attributes is nonlinear.Cross validation and grid search is necessary for parameter search and improving the prediction accuracy.The above approach works well when there are thousands or more data points but for very large data sets a viable approach is to randomly choose a subset of the data set, conduct grid-search on them, and then perform a region-only grid-search on the whole data set.

A graphical user interface (GUI) component is available in the decision support library to allow interactive modification of the rules that have been implemented so far. If necessary, new rule implementations can be created. They are able to use all available patient information when deciding whether he or she is to be included in a training class or not.The library provides implementations of GUI components for displaying DSI trees , data distributions , entry timeline ,and entry details . These are implemented on top of the using Windows Presentation Foundation (WPF).

3.3 Data Access Implementations

There exist two implementations of persistence stores for accessing patient information to be used with the decision support library. One of them uses an entity-attribute-value (EAV) scheme, which is a common methodology for database design in healthcare applications, because of its ability to store heterogeneous and sparse patient data. EAV is befitted for querying data of individual patients but it is found to be inefficient for bulk queries, which are needed for collecting large quantities of training data.

This facilitates the use of a normalized database where the patient and all record types are represented by their own tables. Unfortunately, this is a conflicting requirement for the decision support library, which strives to be a generic one, accepting any kind of data from any clinic to be incorporated into it.A normalized database and persistence store generators have been developed to go along with the library to overcome these conflicting requirements. They are based on the java language features.

4. Results and Discussions

Applicability of the software library is demonstrated by developing a CDSS tool for early prediction of AD and various types of cancers.The complexity of implementation work is evaluated qualitatively and the computational performance of the interactive DSI method is measured quantitatively on a laptop PC with Windows 7, 4 GB of memory, and a 2.4 GHz dual core processor. There are no other CDSS tools or decision support libraries for clinical diagnostics developed with a similar philosophy so far, i.e., using any available sparse and unprocessed patient data, and having the advantage of not requiring manual tuning or

Data Sets	Classification accuracy with SVM Predictor	Classification accuracy with GA-SVM Predictor
Alzheimer Disease	86.56 %	92.45%
Lung Cancer	84.88%	89.94%

decision parameters defined by clinical experts. Features of the library aim to support clinical requirements, e.g.,they accommodate workflows where patient data are collected sporadically.

The DSI method behind the decision support library is able to provide values for quickly interpretable visualizations of multiscale data without compromising the prediction accuracy. Compared to the reference classification methods, the DSI also emphasizes clinical interpretability by 1) providing information about all subsystems of different scales (e.g., genetic, molecular, structural, and neuropsychological) individually and also as a part of the whole, 2) computing a rank of the patient data in relation to diagnosed populations instead of maximizing class separation, which leads to 3) consistency in output that should reflect the magnitude of changes in the raw data.In addition to highlighting important details to clinicians, the DSI and relevance values can facilitate building of expert systems.

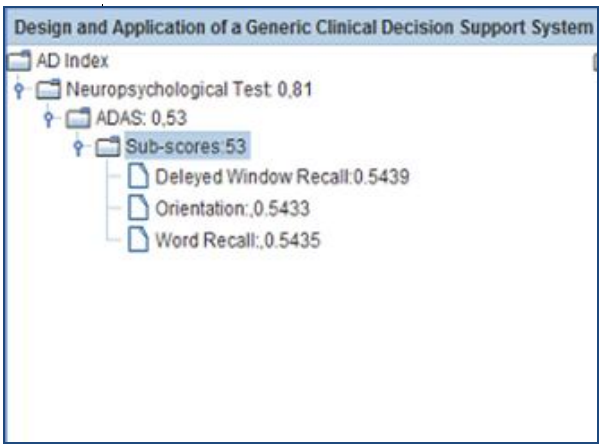

Figure 2: DSI Tree visualization for Alzheimer Disease

With the use of GA-SVM based predictor we are able to achieve more accurate results. To evaluate the classification accuracy of the proposed system in different classification tasks, we tried several realworld datasets from the UCI and ADNI database.

Table 1: Result Summarisation

Healthcare is slowly moving towards electronic healthrecords. Eventually, patient data could be automatically loaded for analyses inside a tool such as this. A clinician diagnosing a patient would not need to observe hundreds of individual measurements at different scales, available from several sources. Instead, they could see all available data at once, hypothesize a disease, and immediately see which data are relevant in that context and which point toward the disease. This could save both time and frustration from information overload.

As a result of our continuing bounded understanding of the complex, dynamic nature of cancers a number of barriers arise to the success of this approach.The other challenges include the often constrained access to appropriate experimental and clinical data; the difficulties in validating models against these data; and the challenges involved in communicating and sharing modeling methods among the field's multiple stakeholders[4].However, by ensuring the collaborative effort and expertise of scientists from different disciplines and on the continuing development of groundbreaking innovative computational and statistical methods,it can be hoped that multiscale cancer modeling will evolve as a motivating domain in guiding targeted experimental research,avoiding adverse events in enabling patient-specific predictions, and thus in accelerating personalized medicine,all achieved at reduced cost and effort[5].

As if now some manual work is needed, for entering patient records into the tool.This limits the presented solution to specialist clinics in the immediate future. Also routinely collected clinical data contain more artifacts and missing information than research data that affect the performance of the methods. Therefore, there are plans for future studies using less well-curated patient data from realistic sources[6].

The main disadvantage of the presented DSI method and the decision support library implementation is that in addition to the patient measurements for analyses, they require properly validated datasets for control and disease cases[7][8]. This training data could be local to a particular clinic, but could also be collected regionally or nationally, greatly decreasing the burden of creating validated training datasets.Data obtained in research studies should be a good starting point for compiling the initial training datasets.Another limitation of the proposed system is that currently the library has proper support for

two-class problems only. Future research will address how these methods are appropriately applied when multiple diseases from different families of diseases are in consideration, which is a clinically important requirement for differential diagnostics[9][10].

5. Conclusion And Future Work

The design and implementation of a generic decision support system is presented. It is implemented as a reusable software library employing a statistical disease state modeling method, which is able to robustly analyze heterogeneous multiscale patient data with minimal preprocessing[11]. The library can be rapidly applied in several contexts.This is attributed to the context-agnostic data access, analysis, and visualization methods. Multi-scale patient-specific modeling is an emerging frontier in computational biology and possibly the future of medicine[12].

When a new problem or data is presented, there is no searching of parameters,development of new user interfaces or handling of missing values. The ultimate objective is to provide evidence-based decision support for clinicians during diagnostic work. Application of the decision support library is demonstrated by developing a prototype CDSS tool for early prediction of various types of cancers and Alzheimer's Disease.The decision support library is applied to several other datasets to assess their robustness more comprehensively.

We conducted experiments to evaluate the classification accuracy of the proposed GA-based approach with RBF kernel.Other kernel parameters can also be optimized using the same approach. The proposed approach can also be applied to support vector regression (SVR).

Acknowledgment

The authors thank participants of project PredictAD, funded partially by the 7th Framework Program by the European Commission under the ICT theme Virtual Physiological Human(Grant Agreement 224328). Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

References

[1] Jussi Mattila, Juha Koikkalainen, Arho Virkki, Mark van Gils, "Design and Application of a Generic Clinical Decision Support System for Multiscale", *IEEE Transactions on Biomedical Engineering*, Vol. 59, NO. 1,2012.

[2] T. S. Deisboeck and G. S. Stamatakos , "Multiscale Cancer Modeling" , *London: CRC Press*,2010.

[3] Cheng-Lung Huang , Chieh-Jen Wang"A GA-based feature selection and parameters optimization for support vector machines",Elsevier Expert Systems with Applications 31 (231–240),2006.

[4] Dean F.Sittig and Adam Wright,"Grand Challenges in Clinical Decision Support.", *Journal of Biomedical Informatics* 41,2008.

[5] Joan S Ash and Dean F Sittig.,"Recommended practices for computerised clinical decision support and knowledge management in community settings: a qualitative study.", *BMC*,2012.

[6] Bonnie Kaplan,"Evaluating informatics applications—m,clinical decision support systems literature review" ,*International Journal of Medical Informatics* 64, 15–37,2001.

[7] Kuan-Liang Kuo & Chiou-Shann Fuh, ,"A Rule-Based Clinical Decision Model to Support Interpretation of Multiple Data in Health Examinations",*Springer Science+Business Media*, LLC,2009.

[8] J. Mattila, J. Koikkalainen, A. Virkki, A. Simonsen, M.van Gils, G.Waldemar ,H.Soininen, and J.Lotjonen , "A disease state fingerprint for evaluation of alzheimer's diseases," *The J. Alzheimer'sDis.*,[Online].Available:http://iospress.metapress.com/content/kg543256311310.

[9]R. A. Greenes Ed., "Clinical Decision Support the RoadAhead",*New York: Elsevier*,2007.

[10]Emanuelen Lettierie ,Giovanni Radaeli,Christina Masella, "Information Systems and Change Management in Healthcare:the (un)solved quest for changing physician's behaviour",*Int.J.Information Systems and Change Management*,Vol 4,No 3,2010.

[11] H. Chen, S. Fuller, C. Friedman, and W. Hersh, *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*. NewYork: Springer, 2010.

[12]J. B. Bassingthwaighte and H. J. Chizeck, "The physiome projects and multiscale modeling," *IEEE Signal Processing Magazine*, 2008.

[13] E. Alpaydin, *Introduction to Machine Learning*, 2nd edition. Cambridge:MIT Press, 2009.

[14] Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining: Concepts and Techniques*, 3rd edition, Morgan Kaufmann, 2011.

[15]Ian H Witten and Eibe Frank,Data Mining Practical Machine Learning Tools and Techniques,2nd edition,2005.

[16] "Physiome project, http://www.physiome.org/."

[17]A. Frank and A. Asuncion,UCI Machine Learning Repository,[Online]Available:http://archive.ics.uci.edu/ml, 2010.