INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

DATA PULL OUT AND FACTS UNEARTHING IN BIOLOGICAL DATABASES

Dr. Prasadu Peddi
Assistant Professor, Dept of CSE Shri Jagdishprasad Jhabarmal Tibrewala university, Rajasthan.

Abstract: *New technologies for Knowledge Discovery from Databases and Data Mining promise new insights into an ever-growing volume of biological data. KDD technology can be used in conjunction with laboratory experimentation to speed up biological research. This article provides an overview of KDD and a discussion on data mining tools and their biological applications. We will discuss domain concepts related to databases and biological data, as well as the latest KDD and data mining developments within biology.*

1 INTRODUCTION

The number of biological databases continues to increase rapidly. This rapid growth is reflected in an increase in the complexity and size of individual databases, as well as the proliferation of new databases. This vast amount of data allows for the extraction and interrelationship of high-level information. It also makes it possible to discover interesting patterns in these databases. KDD, a new field that combines techniques from statistics, articial intelligence and databases, is concerned with the theoretical as well as practical aspects of extracting high-level information (or knowledge), from large volumes of low-level data. Low-level data can be used to extract high-level information. These forms include shorter reports, more abstract models (e.g. descriptive models of data generation), and more useful (e.g. redictive models for estimating future cases). Fayyad and colleagues. Fayyad et al. (1996) define KDD as the entire process of discovering useful information from databases. Data mining is a specific step in that process. Knowledge discovery in databases is the non-trivial process that identifies valid, novel, potentially valuable, and understandable patterns within data. A data set is a collection of facts stored in a file. A pattern is an expression in some language that describes a subset or applies a model to that subset. A pattern can also be extracted by:

(a) fitting data to a model.

(b) Finding structure from data

(c) Providing a high-level description for a data set. KDD is interactive and iterative. KDD has multiple steps that include:

(a) Data preparation

(b) Pattern searching

(c) Knowledge evaluation

(d) refinement.

These steps can be repeated in

Multiple iterations. KDD's core activity is data mining + the use of specific tools to extract and discover patterns. KDD uses search and inference methods instead of simple calculations.

Validity of the discovered patterns means that the user must be able to predict the correct results from the new data. There are many ways to

ISSN: 2057-5688

measure validity, such as prediction accuracy using new data or utility or gain (e.g. speed-up or dollar value). It is subjective to assess the novelty, usefulness or understanding of the newly discovered knowledge. This depends on the KDD. The overall measure of a pattern's value is called interestingness. It includes validity, novelty and usefulness. Biases and errors are inherent in biological data. KDD can be improved by filtering out errors and de-biasing the data. Filtering errors and de-biasing data can be done at any stage of the KDD process. It is often performed based on human judgment. Alternately, filtering can be done within the data mining algorithm. Validation of the newly discovered knowledge is critical for both the data mining and the overall KDD process. Verification tasks are a form validation. They involve the estimation of the quality and use statistical tests. Validation is essential for discovery, and especially prediction tasks. In designing a KDD process for biology, it is important to consider both the requirements of the KDD process as well as the requirements specific to the application domain. In the study of biological sequences, data mining tasks were defined. Examples include the discovery of genes in DNA sequences and regulatory elements within genomes (e.g. Brazma, 1997)

Knowledge discovery and transmembrane domain research (Shoudai and et.al., 1995). There are many tools that can be used for data mining in biology, but it is not easy to choose the right tool. KDD allows for the selection of appropriate data mining methods, taking into consideration both domain characteristics as well as general KDD processes.

2) Introduction to KDD

Data is raw material that must be processed by humans, computers, or any other means. Information is data that is organized by a person or a machine to make it meaningful and useable. Conventional databases are simple data types like numbers, strings, and Boolean value. Knowledge can be described as information that includes raw data, interpreted data, and expertise. Modern applications require more complex information, such as procedures, actions and causality. Information on structure and organization is also required for biological applications. This is the larger category of information known as knowledge.

KDD involves ten steps. Fayyad et al. defined the first nine. (1996).

Step by Stranieri and Zeleznikov (personal communication).

1. Learn the application domain +- This includes developing relevant prior knowledge

From the perspective of the user, identify the goal and initial purpose of the KDD process.

2. Create a target data collection +- this includes selecting a data set, focusing on one set of variables, or data

Samples on which discovery is to take place

Data cleaning and pre-processing +- involves operations like removing noise and outliers, gathering the information necessary to model noise and making decisions about how to handle missing data fields.

Data reduction and projection +- involves finding useful features to represent data. The number of variables that are being considered can be

ISSN: 2057-5688

decreased or transformed to make it more consistent.

5. The function of data mining +- is determining the purpose of the model generated by the data mining algorithm: classification, regression, and clustering.

6. Selecting the data mining algorithms +- involves selecting the methods that can be used to search for patterns in data and match a particular data mining technique with the overall criteria for the KDD process. This includes selecting the right parameters and models. This involves comparing a data mining method to the KDD.

7. Data mining +- is the search for patterns of interest in particular representational forms or a group of representations, including regression rules or trees, clustering, dependency modeling, and classification rules.

8. Interpretation +- can be used to indicate that there are additional iterations of any step (1+7). This step can also be

Visualisation of extracted patterns and models, or visualisations of data from the models extracted;

This step is called "Using discovered knowledge +". It involves taking immediate action on the newly discovered knowledge, including incorporating it into another system or reporting the knowledge. This includes resolving any potential conflicts with previously extracted knowledge.

Evaluation of KDD purpose +- Newly discovered knowledge can often be used to form new hypotheses. Also, new questions could be raised with the expanded knowledge base. This step allows the KDD process to be evaluated in order to refine and expand its purpose relative the previous KDD cycle. Diagrammatic representation of KDD

2.1 Data mining:

Data mining is a problem-solving methodology that finds a formal description, eventually of a complex nature, of patterns and regularities in a set of data. Decker and Focardi (1995) consider various domains that are suitable for data mining, including medicine and business. They state that in practical applications, data mining is based on two assumptions. First, the functions that one wants to generalize can be approximated through some relatively simple computational model with a certain level of precision. Second, the sample data set contains suQient information required for performing the generalisation. The additional steps in the KDD process are essential to ensure that useful knowledge is derived from the data. Blind application of data mining known as data dredging can easily lead to the discovery of meaningless or misleading patterns.

2.2 Data mining tools: an overview

We shall briefly describe several popular techniques, namely: (a) decision trees and rules, (b) nonlinear regression and classification methods, (c) example-based methods, (d) probabilistic models, and (e) relational learning models. In this paper, we provide an overview of data mining methods intended to help the reader understand the data mining methods and facilitate the selection of the most appropriate method for a given problem.

3 Research Methods

Decision trees and rules:

ISSN: 2057-5688

Decision trees consist of nodes and edges; each node contains a test on some attribute of the data. Decision trees and rules that use binary splits produce classifications which can be easily understood and produce compact models. Hhowever, the restriction to a particular tree or rule

representation can limit the functionality and approximation power of the model. An example of a decision tree is given in Figure 1. Decision trees use likelihood-based model-evaluation methods

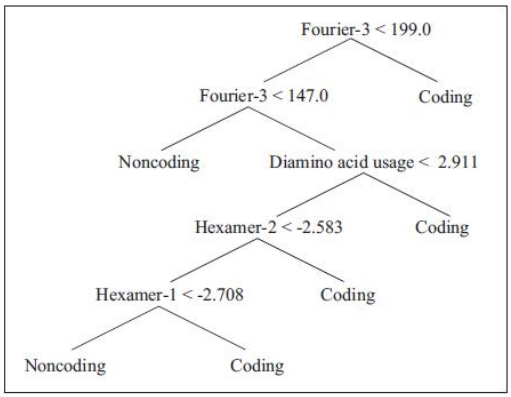


Figure 1 An example of a decision tree which predicts protein coding regions. This decision tree uses four features and contains +ve test nodes.

combined with search methods for growing and pruning tree structures. Decision trees and rules are commonly used in prediction tasks for classification, regression or summarization. Induction is the process in which rules are generated from sample cases. A rule induction system creates rules that fit the example cases. The rules can be used to assess other cases where the outcome is not known. An example of a rule induction system is the ID3 algorithm of Quinlan (1986), which has been extended to the C4.5 algorithm (Quinlan, 1993), and more recently to C5.0. A characteristic of induction algorithms is that the learning is based on the statistical analysis of the training set. Machine induction allows for the deduction of new knowledge. It may be possible to list all the factors in influencing a decision without understanding their impacts. The rules generated can be reviewed and modified by the domain expert.

Nonlinear regression and classification methods: Non-linear regression methods utilise non-linear functions such as polynomials, sigmoids, or splines, for finding relationships between input variables Xi and output variables Yi, by fitting functions to the available data. Examples include methods which use (a) feedforward neural networks, (b) adaptive splines, or (c) projection pursuit. Non-linear regression methods, although powerful in representational power, can be difficult to interpret. Non-linear regression methods utilize non-linear functions such as polynomials, sigmoids, or splines, for finding relationships between input variables Xi and output variables Yi, by fitting functions to the available data. Examples include methods which use: (a) feedforward neural networks, (b) adaptive splines, or (c) projection pursuit. Non-linear regression methods, although powerful in representational power, can be difficult to interpret. A neural network of the appropriate size can universally approximate any smooth function to any desired degree of accuracy. However, it is relatively difficult to elucidate generalized rules that

characterize training data from a trained neural network. Artificial neural networks were originally

ISSN: 2057-5688

designed to simulate the information processing (connectivity and signalling) within a biological brain. This consists of many self-adjusting processing elements cooperating in a densely interconnected network.

A Hidden Markov model (HMM) is a class of probabilistic graphical models. It is defined by a finite set of states, associated with a (usually multidimensional) probability distribution. Transitions between the states are governed by a set of transition and emission probabilities. An outcome of a transition from a particular state can be generated, according to the associated probability distribution. The states are not visible to an external observer, and therefore they are hidden to the outside; only the outcome is visible. The assumption in first-order Markov model is that the transitions depend only upon the current state. HMMs can be trained using sets of pre-classified examples and a variety of learning algorithms. The advantages of HMMs are that they combine solid statistical basis with eQient learning algorithms. The limitations of HMMs include the need for a large number of free parameters, which in turn require a significant number of training cases. Further, a good knowledge of the domain model is required for selecting the appropriate HMM architecture for a specific task. HMMs have been extensively used in modelling biological sequence data.

Table: Definition of terms for assessing the accuracy of predictive models

	Experimental positives	Experimental negatives
Predicted positives	True positives (TP)	False positives (FP)
Predicted negatives	False negatives (FN)	True negatives (TN)

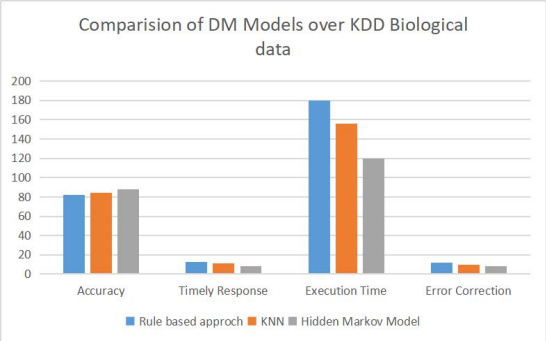
Accuracy measure	Formula	Pairs with
Sensitivity	SE = TP/(TP+FN)	SP
Specificity	SP= TN/(TN+FP)	SE
Positive predictive value	PPV = TP/(TP+FP)	NPV
Negative predictive value	NPV = TN/(TN+FN)	PPV
Accuracy	Acc = (TP+TN)/(TP+TN+FP+FN)	-
Aroc	Integration of ROC curves	-

Acc and Aroc provide the convenience of a single measure of the accuracy of predictive models. The Acc measure is suitable when prevalence of positive and negative cases is similar and therefore is often not useful in prediction of biological eQects. Using Relative Operation Characteristics (ROC) for the integration of functions of (1-SP, SE) for various decision thresholds provides the Aroc measure. Values of Aroc = 50% indicate random-choice, Aroc >80% good accuracy, and Aroc >90% excellent accuracy of predictions. A variety of theoretical methods exist including splitting data into training and test sets, internal crossvalidation and bootstrapping. Theoretical estimates of accuracy tend to be somewhat

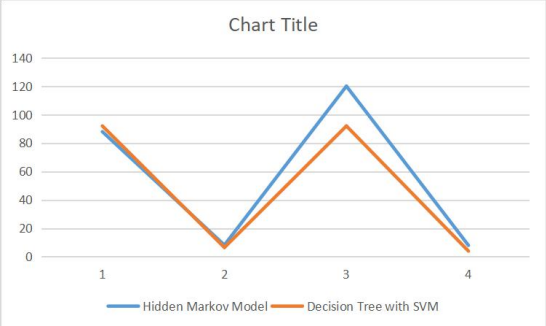
ISSN: 2057-5688

optimistic. Experimental testing of theoretical models is the best validation option, provided that the experimental method is of an acceptable accuracy.

4 Results and Discussion



Comparison of models for biological kdd databases



HMM and DT model comparison for accuracy

5 Conclusion

Automated methods of generating biological data are a recent development in biology. Bioinformatics will face a significant challenge due to the complexity, growth, and increasing amount of genomic data. This has created the need for technologies to support data handling and interpretation. Data accumulation has slowed down significantly in terms of progress in automatic data handling. This disparity has many consequences, including the persistence and spread of incorrect information as well as scientific insights being overlooked. KDD technology allows for data extraction and automation, as well as support for the interpretation of that knowledge. Another problem that results from data accumulation is the difficulty in planning and selecting wet-lab experiments. Computer models can be used to supplement laboratory experiments and accelerate the KDD process in biology. They showed that large-scale experiments can be avoided by using smaller, targeted experiments. These experiments were designed to validate and develop appropriate computer models. These models can be used to perform large-scale computer-simulated experiment quickly and cheaply. Computer models will become increasingly important in biology research. KDD technology allows for the efficient and complete use of computer models in biology research.

6 Future Scope:

Process the genome data with the KDD Data mining methods To study the pattern matching and clustering models in the data generated from the KDD datasets Deploying the Machine learning models over the data and implement the nerual networks concpets to analyse the data with better accruay and perfomrance.

7 References:

[1] Mount, D. W., "Bioinformatics: Sequence and genome analysis". Spring Harbor Press, (2002).

[2] Cristianini, N. and Hahn M., "Introduction to computational genomics", Cambridge University Press, 2006. ISBN 0-52167191-4.

[3] Sindhu, S. and Sindhu, D., "Development of computational tools for metabolic engineering".

International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, issue 5, pp. 9208-9217, 2016a.

[4] Li, J., Wong, L. and Yang, Q., "Data mining in bioinformatics", IEEE Intelligent System, IEEE Computer Society, 2005.

[5] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R., "Advances in knowledge discovery and data mining". AAAI Press/MIT Press, Menlo Park, California, USA, 1996.

[6] Sindhu, D. and Sindhu, S., "Biological computers: Their application in gene mining and protein engineering", International Journal of Technical Research, Vol. 4, issue 3, pp. 15-21, 2015.

[7] Luscombe, N.M., Greenbaum, D. and Gerstein, M., "What is bioinformatics? A proposed definition and overview of the field". Methods of Information in Medicine, Vol. 40, issue 4, pp. 346-358, 2001.

[8] Elizabeth P., Isaac L. and Kevin T., "Computational modeling approaches for studying of synthetic biological networks", Current Bioinformatics, Vol. 3, pp. 1-12, 2008.

[9] Sindhu, D. and Sindhu, S., "Computational programming for product designing in synthetic biology", International Journal of Innovative Rsearch in Science, Engineering and Technology, Vol. 5, issue 5, pp. 8095-8103, 2016b.

[10] Prasadu Peddi (2018), Data sharing Privacy in Mobile cloud using AES, ISSN 2319-1953, volume 7, issue 4.

[11] Cameron D.E., Bashor C.J. and Collins J.J., "A brief history of synthetic biology", Natural Review Microbiology, Vol. 12, pp. 381-390, 2014.

[12] Chandran D., Bergmann F.T. and Sauro, H.M., "Tinker Cell: modular CAD tool for synthetic biology, Journal Biological Engineering, Vol. 3, pp. 19-36, 2009.

[13] Purnick P.E. and Weiss R., "The second wave of synthetic biology: from modules to systems. Natural Review of Molecular and Cellular Biology, Vol. 10, pp. 410-422, 2009.

[14] Chadha, P. and Singh, G.N., "Classification rules and genetic algorithms in data mining", Global Journal of Computer Science and Technology Software and Data Engineering, Vol. 12, issue 15, pp. 1-5, 2012.

[15] Han, J., "How can data mining help bio-data analysis"? In: Zaki, M.J., Wang, J.T.L. and Toivonen, H.T.T. (Eds). Proceedings of the 2nd ACM SIGKDD Workshop on data mining in bioinformatics, Vol. 1-2, 2002.

[16] Prasadu Peddi (2017) "Design of Simulators for Job Group Resource Allocation Scheduling In Grid and Cloud Computing Environments", ISSN: 2319- 8753 volume 6 issue 8 pp: 17805-17811.

[17] Molla, M., Waddell, M., Page, D. and Shavlik, J., "Using machine learning to design and interpret gene-expression microarrays", AI Magazine, Vol. 25, issue 1, pp. 23-44, 2004.

[18] Krieger, M., Scott, M.P, Matsudaira, P.T., Lodish, H.F., Darnell, J.E., Lawrence, Z., Kaiser, C. and Berk, A., "Structure of nucleic acids, Section 4.1", Molecular Cell Biology. New York: W.H. Freeman and Co. 2004.

[19] Jacob, F. and Monod, J., "Genetic regulatory mechanisms in the synthesis of proteins", Journal of Molecular Biology, Vol. 3, pp. 318–356, 1961.

ISSN: 2057-5688

[20] Prasadu Peddi (2016), Comparative study on cloud optimized resource and prediction using machine learning algorithm, ISSN: 2455-6300, volume 1, issue 3, pp: 88-94