Tree Mining in Mental Health Domain

Maja Hadzic, Fedja Hadzic, Tharam Dillon
Digital Ecosystems and Business Intelligence Institute
Curtin University of Technology
GPO Box U1987, Perth 6845
Western Australia, Australia
E-mail: m.hadzic@curtin.edu.au, fedja.hadzic@postgrad.curtin.edu.au, t.dillon@curtin.edu.au

Abstract

The number of mentally ill people is increasing globally each year. Despite major medical advances, the identification of genetic and environmental factors responsible for mental illnesses still remains unsolved and is therefore a very active research focus today.

Semi-structured data structure is predominantly used to enable the meaningful representations of the available mental health knowledge. Data mining techniques can be used to efficiently analyze these semi-structured mental health data. Tree mining algorithms can efficiently extract frequent substructures from semi-structured knowledge representation such as XML.

In this paper we demonstrate effective application of the tree mining algorithms on records of mentally ill patients. The extracted data patterns can provide useful information to help in prevention of mental illness and assist in delivery of effective and efficient mental health services.

1. Introduction

The World Health Organization predicted that depression would be the world's leading cause of disability by 2020 [1]. The recognition that mental health is costly and many cases will not become chronic if treated early has lead to an increase in research in the last 20 years. The research covers different mental illnesses with a huge range of results regarding different disease types, symptoms, treatments, disease causing factors (genetic and environmental) as well as candidate genes that could be responsible for the onset of these diseases. Usually in any medical research, one research team examines one cause, be it genetic or environmental, for one type of disorder. But for mental illnesses, in order for research to be inclusive, we need to look at all factors simultaneously. We need a tool that combines and examines all the genetic factors together, all the environmental factors together and all genetic and environmental factors together. Also, a huge body of patient's information is available.

All biological experiments are driven by a plethora of experimental design hypotheses to be proven or rejected based on data values stored in multiple distributed biomedical databases, for example, genome or proteome databases. To extract and analyze the data poses a much bigger challenge for researchers than to generate the data [2]. We need to take a systematic approach to making use of this information. The true value of this information can be significantly increased through smart information processing and analysis. There is a crying need for information technologies to be implemented within mental health and health domain in general. In their paper on the general health need of adults with severe mental illness Horvitz-Lennon *et al.* [3] state that we need to fully embrace information technology and its potential for improving service efficiency and develop a better information infrastructure for the patient's care. Frequent pattern analysis has been a focused theme of study in data mining, and many algorithms and methods have been developed for mining frequent sequential and structural patterns [4, 5, 6].

We believe that data mining approach can significantly help the research into mental illness. Data mining helps find patterns and knowledge that are embedded in the data. It requires exploration and analysis of large quantities of target data for the purpose of better understanding and deriving knowledge regarding the problem at hand. Because mental illnesses are caused by a number of different genetic and environmental factors, data mining technique would be the best approach to derive relationships between the illness-causing factors and specific illness type. Data mining can be applied on the knowledge available via published information or on patient's data or on both, general knowledge and patient's data.

Tree Mining has attracted lots of interest among the data mining community, due to the increasing use of semi-structured data sources for more meaningful knowledge representations. This is particularly evident in areas such as Bioinformatics, XML Mining, Web applications, scientific data management, and more generally in any area where the knowledge is represented in a tree-structured form. Many powerful tree mining algorithms

have been developed to aid in structural comparisons, association rule discovery and in general, for mining of tree structured knowledge representations.

The problem of frequent subtree mining can be generally stated as follows. Given a tree database Tdb and minimum support threshold (σ), find all subtrees that occur at least σ times in Tdb. Within this framework the two most commonly mined types of subtrees are induced and embedded. An induced subtree preserves the parent-child relationships of each node in the original tree. In addition to this, an embedded subtree allows a parent in the subtree to be an ancestor in the original tree and hence ancestor-descendant relationships are preserved over several levels. The subtrees can be further distinguished based upon the ordering of siblings. An ordered subtree preserves the left-to-right ordering among the sibling nodes in the original tree while in an unordered subtree this ordering is not preserved.

Our work in the area of frequent subtree mining is characterized by adopting a Tree Model Guided (TMG) [7,8] candidate generation as opposed to the join approach [24] which is commonly used. This non-redundant systematic enumeration model ensures only valid candidates are generated which conform to the underlying tree structure of the data. Furthermore, our unique Embedding List [7] representation of the tree structure has allowed for an efficient implementation of the TMG approach which resulted in efficient algorithms MB3-Miner [7] and IMB3-Miner for mining of ordered embedded and induced subtrees, respectively. Razor algorithm [9] was a further extension developed for mining embedded subtrees where the distance of nodes relative to the root of the subtree needs to be considered. With respect to unordered subtree mining we have developed the UNI3 [10] algorithm for mining of induced unordered subtrees.. We have previously applied our algorithms on large and complex tree structures and experimentally demonstrated their scalability [7, 8].

From the application perspective, in [11] we have applied our tree mining algorithm for the analysis of Protein Ontology [12] database for Human Prion proteins which was represented in XML format. We have shown how tree mining aids in discovering of useful pattern structures in Protein Ontology datasets. This makes it useful for comparison of protein datasets taken across protein families and species and helps in discovering interesting similarities and differences.

Our main aim in this paper is to demonstrate the potential of the tree mining algorithms in deriving useful knowledge patterns in metal health domain. In this paper, we make use of our previously developed IMB3-Miner [8] algorithm to show how tree mining techniques can be applied on patient's data represented in XML format. The IMB3-Miner algorithm is well scalable when applied to large datasets consisting of complex structures, as was experimentally demonstrated in [8]. However, in order to make the results clearer and the underlying principle easier to understand, simpler synthetic datasets were used. The implications of using different mining parameters within the current tree mining framework are discussed, and the potential of extracted patterns in providing useful information is demonstrated.

The rest of the paper is organized as follows. In Section 2, we discuss major issues associated with study and control of mental illnesses, and give an overview of existing work in tree mining domain. In Section 3, basic tree mining concepts are defined and illustrated. An overview of the five steps associated with mining of patient's records is given in Section 4. In Section 5, the data mining approach is illustrated on an artificial dataset. The significance and benefits of this research is discussed in Section 6. Our final conclusion is given in Section 7.

2. Related Work

Human diseases can be described as genetically simple (Mendelian) or complex (multifactorial). Simple diseases manifest the presence of a mutated gene and the disease. Mental illnesses do not follow Mendelian patterns but are caused by a number of genes usually interacting with various environmental factors [13]. Some mental illnesses may be caused by a physical dysfunction of the brain but it is not known what triggers this exactly and the exact cause of many mental illnesses are unclear. For example, genetic analysis has identified candidate loci on human chromosomes 4, 7, 8, 9, 10, 13, 14 and 17 [14]. There is some evidence that environmental factors such as stress, life-cycle matters, social environment, climate etc. are important [3, 15]. The situation of mental illnesses is made even more complicated through the existence of different illness types, such as depression, bipolar disorder and schizophrenia, and illness subtypes, such as bipolar I and II. Despite major medical advances, the identification of genetic and environmental factors responsible for mental illnesses still remains unsolved and is therefore a very active research focus today.

Data mining is a set of processes that is based on automated searching for actionable knowledge buried within a huge body of data to extract information and find hidden patterns and behaviors for the purpose of making predictive models for decision making and new discoveries. Within the biomedical and health field, data mining techniques have been predominately used for tasks such as text mining, gene expression analysis, drug design, genomics and proteomics [16]. The data analysis necessary for microarrays has necessitated data mining [17]. Recently, use of data mining methods has been proposed for the purpose of mapping and identification of

complex disease loci [18]. However, the proposed methods are yet to be implemented.

Tree mining algorithms are increasingly being developed and the scope of their application usually depends on the assumptions made about the tree structure that the algorithm can be applied to. Naturally, these assumptions depend upon the domain of interest, where the developed algorithm is to be applied. Hence, many tree mining algorithms exist and they can be distinguished based upon the types of tree patterns that they extract. PathJoin [19], uNot [20], uFreqt [21], and HybridTreeMiner [22], mine induced, unordered trees. AMOT [23], mines induced ordered trees, and by using 'right-and-left tree join' method it efficiently enumerates only those candidates that have a high probability of being frequent. Treeminer [24], is an efficient algorithm for discovering all frequent embedded subtrees in a forest using a data structure called the vertical scope-list. The idea was extended to the SLEUTH [25] algorithm for mining frequent embedded unordered subtrees. TreeFinder [26] uses an Inductive Logic Programming approach to mine unordered, embedded subtrees, but in the process many frequent subtrees can be missed. In regards to the application of tree mining to biological data, some approaches have been developed for analysis of phylogenetic databases [27, 28].

3. Basic tree mining concepts

This section provides more formal definitions of some basic tree concepts accompanied with some illustrative examples. Only concepts necessary for understanding the current work are discussed. For a more extensive overview of the area including various implementation issues and algorithm comparisons we refer the interested reader to [7, 8, 10].

The problem of frequent subtree mining can be generally stated as: given a tree database *Tdb* and minimum support threshold (σ), find all subtrees that occur at least σ times in *Tdb*.

3. 1 Trees and subtrees

The underlying structure of an XML document can be viewed as a tree. A patient record can be captured by a XML document analogous to the one shown in Figure 1. On this running example we will illustrate application of tree mining technology within mental health domain. We aim to capture the information about causal factors of severe mental illnesses (depression, bipolar disorder and schizophrenia) [3]. Genetical causes include the presence of a mutated gene in patient's DNA (represented by Gene a, Gene b and Gene c) while environmental factors cover social, economic, physical and cultural aspects (such as climate, drugs misuse, family conditions, economic conditions and stress). All this causal factor together play an important role on the onset of mental illnesses.

A tree can be denoted as $T(v0,V,L,E)$, where:
(1) $v0 \in V$ is the *root* vertex;
(2) V is the set of *vertices* or *nodes*;
(3) L is the set of *labels* of vertices, for any vertex $v \in V$, $L(v)$ denotes the label of v; and
(4) $E = \{(x,y) | x,y \in V\}$ is the set of *edges* in the tree.

```
<?xml version="1.0" ?>
<element name = "Type">
<element name = "Cause">

   <element name="Genetic">
         <element name="Gene a" />
         <element name="Gene b" />
         <element name="Gene c" />
   </element>

   <element name="Environmental">
         <element name ="Climate" />
         <element name ="Drugs misuse" />
         <element name ="Family conditions" />
         <element name ="Economic conditions" />
         <element name ="Stress" />
    - </element>

</element>
</element>
```

Figure 1 XML representation of patient's records

A root is the topmost node in the tree. The Parent of node v is defined as the predecessor of node v. A node without any child is a leaf node; otherwise, it is an internal node. If for each internal node, all the children are ordered, then the tree is an *ordered tree*.

In Figure 2, we represent the XML document from in Figure 1 as a tree.

Some properties of the tree represented in Figure 2, are:

- 'type' $\in V$ is the *root* vertex
- L = {type, cause, genetic, Gene a, Gene b, Gene c, environmental, climate, drugs misuse, family conditions, economic conditions, stress}
- The *parent* of node 'cause' is node 'type', of node 'genetic' is node 'cause', of node 'Gene a' is node 'genetic' etc.
- 'Gene a', 'Gene b', 'Gene c', 'stress', 'drugs misuse', 'family conditions', 'economic conditions' and 'stress' are *leaf* nodes.
- 'Genetic', 'environmental' and 'cause' are *internal nodes*.

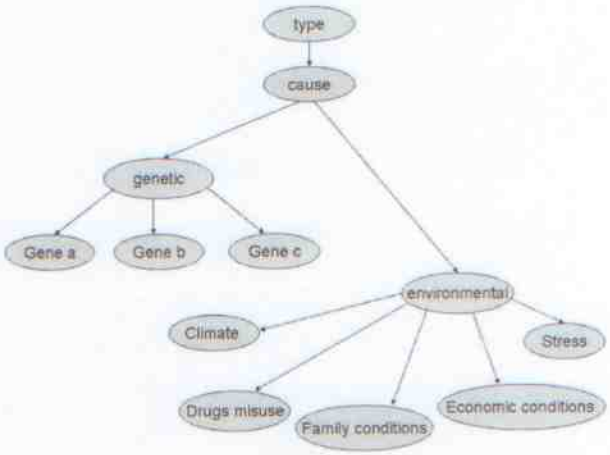

Figure 2 Patient record represented in a tree structure

Within this framework the two most commonly mined types of subtrees are induced and embedded. An induced subtree preserves the parent-child relationships of each node in the original tree while an embedded subtree allows a parent in the subtree to be an ancestor in the original tree. Formal definitions follow:

(1) **Induced** subtree
A tree $T'(r', V', L', E')$ is an induced subtree of a tree $T(r, V, L, E)$ iff (1) $V' \subseteq V$, (2) $E' \subseteq E$, (3) $L' \subseteq L$ and $L'(v)=L(v)$, (4) $\forall v' \in V'$, $\forall v \in V$, v' is not the root node, and v' has a parent in T', then parent(v')=parent(v)

(2) **Embedded** subtree.
A tree $T'(r', V', L', E')$ is an *embedded subtree* of a tree $T(r, V, L, E)$ iff (1) $V' \subseteq V$, (2) if $(v_1, v_2) \in E'$ then $parent(v_2) = v_1$ in T', only if v_1 is ancestor of v_2 in T and (3) $L' \subseteq L$ and $L'(v)=L(v)$.

In addition to the previous definitions, the subtrees can be further distinguished based upon the ordering of siblings:
(1) **Ordered** subtree, or
It preserves the left-to-right ordering among the sibling nodes in the original tree.

(2) **Unordered** subtree.
For an unordered subtree the order of the siblings (and the subtrees rooted at sibling nodes) can be exchanged and the resulting subtree would be considered the same. Examples of different subtree types related to the Figure 2 are given in Figure 3 below. Please note that induced subtrees are also embedded.

3. 2 Support definitions

In data mining field a transaction has been defined as a set of one or more items obtained from a finite item domain, and a dataset is a collection of transactions [29]. Hence, in context of a tree database, a transaction would correspond to a fragment of the database tree whereby an independent instance is described. In our example, a transaction corresponds to a tree-structured patient record.

Within the current tree mining framework, the support definitions available are transactional support, occurrence-match, and hybrid support [24].

(1) Transactional support checks for the existence of items in a transaction. The support of an item equals to the number of transactions where the item exists. For example, the support of item 'stress' is equal to the number of transactions (i.e. patient records) where it occurs.

(2) Occurrence-match support takes into account repetition of items within a transaction and counts the total occurrences in the dataset as a whole. For example, the support of item 'stress' is equal to the total number of occurrences of item 'stress' within the whole dataset.

(3) Hybrid support is a combination of both whereby extra information about the intra-transactional occurrences of a subtree is provided. Using hybrid support threshold of $x|y$, a subtree is considered frequent if it occurs in x transactions and it occurs at least y times in each of those x transactions. For example, using hybrid support threshold of 10|7 means that a subtree 'depression' → 'cause' → 'environmental' → 'stress' occurs at least 7 times in 10 transactions.

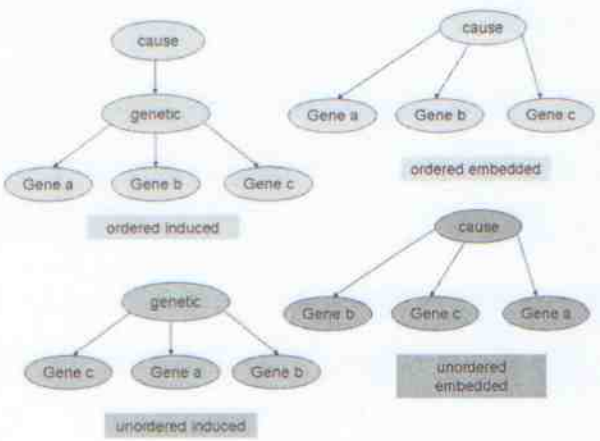

Figure 3 Example of different subtree types

4. Tree mining of patient's records

In order to perform an effective tree mining following steps need to be carried out:

1. Data selection and cleaning
2. Data formatting
3. Tree mining
4. Pattern discovery
5. Knowledge testing and evaluation

In the first phase, we focus on data selection and cleaning. Most of the databases also contain information that is not needed by a data mining application, e.g. a database may also contain patient's records of other illnesses we are not interested in. For this reason, a data set which contains only relevant information needs to be selected. Furthermore, noise and inconsistent data need to be removed from the selected data set. Data selection and cleaning needs to be done carefully; some data may appear to be irrelevant and removed by mistake from the data set, but in fact represents true exceptional cases.

In the second phase, we focus on data formatting. In order to simultaneously analyze all the data within this given data set, the following two conditions need to be fulfilled:

(a) all the data need to be put into the same format

(b) the chosen format needs to be understandable by the data mining application

If there is only access to relational type of data and a set of features can be grouped according to some criteria, XML can be used to represent the relationships between data objects in a more meaningful way. Hence, the relational data could be converted into XML format.

When using tree mining algorithms one has to consider what particular type of subtree is most suitable for the application at hand. Patient's records are most likely to be stored in the same format and attribute values will be ordered the same way among the collected records.

The format and ordering of the data coming from one organization is expected to be the same. In respect to the current tree mining framework this means that mining of ordered subtrees will be most suitable. However, if the collected data originate from separate organizations and these data are found in different formats, then unordered trees would be more suitable. The reason is that the subtrees where the order of sibling nodes is different would still be grouped to the same candidate, and hence the common characteristics of a particular illness would still be found.

Another choice to be made is whether we should be mining induced or embedded subtrees. Since the aim is to extract the patterns where particular patient information has to stay in the context where it occurred, the relationship of nodes in the extracted subtrees should be limited to parent-child relationships. By allowing ancestor-descendant relationships some information could be lost about the context where particular disease characteristic occurred. This is mainly due to the fact that some features of the dataset may have a similar set of values and hence indicating which value belonged to which particular feature is necessary. Therefore we will use our IMB3-Miner algorithm for mining induced ordered subtrees from a database of rooted labeled ordered subtrees (XML).

One further choice to be made is which type of support definition should be used, which is dependant on how the data is organized. Consider the following three scenarios.

Scenario 1: Each patient record is stored as a separate subtree or transaction in the XML document and records describing different illnesses are stored in separate documents. In our example, this would mean that there are three different XML documents each containing same number of transactions as there are patients associated with this illness type (depression, bipolar disorder and schizophrenia). In this case both, occurrence match or transaction based support would be suitable.

Scenario2: Each patient record is stored as a separate subtree or transaction in the XML document but now one XML document contains patient records for all illness types. In our example, this would mean that there is one XML document where the number of transactions is equal to the number of patient records. Here the transactional support would be more appropriate.

Scenario 3: The XML document is organized in such way so that a collection of patient records for one particular illness is contained in one transaction. In our example, this would mean that there is one XML document with three transactions each corresponding with specific illness types. Each of those three transactions would contain records of patients associated with that particular illness type. Hybrid support definition is most suitable in this case.

For scenarios 1 and 2 the minimum support threshold should be chosen to be approximately close to the amount of patient records that the dataset contains about a particular illness in order to find the common characteristic that occurs with all patients having that illness. However, due to noise often being present in data the support can be set lower but not too low as then irrelevant factors may be picked up as important. For scenario 3 the number of illnesses described would be used as the transactional part of the hybrid support, while the approximate number of patient records would be used as the requirement for occurrence of a subtree within each transaction. In this paper we are concerned with the second scenario.

Before data mining takes place, the data set can be split into two subsets, one for deriving the knowledge model ('internal data' from Figure 4) and one for testing the derived knowledge model ('external data' from Figure 4). External data can also come from another organization.

During pattern discovery phase, precise combinations of genetic and environmental illness-causing factors associated with each type of the three mental disorders are identified. The results that establish the interdependence between genetic and environmental factors would make a breakthrough in the research, control and prevention of mental illnesses.

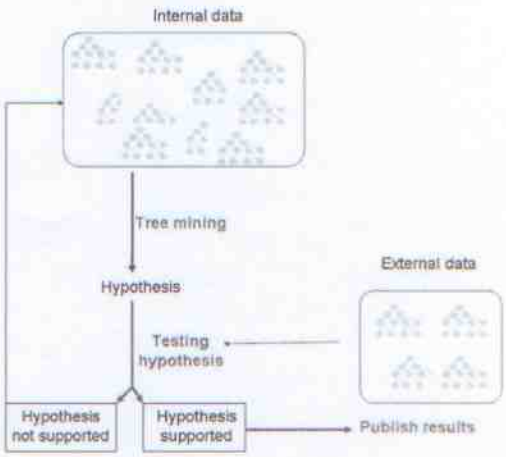

Figure 4 Testing and evaluation of the derived knowledge

Testing and evaluating of the knowledge derived through the data mining applications is illustrated in Figure 4. The 'internal data' correspond to the subset of the data used to derive the hypothesis while the 'external data' correspond to the subset of data 'unseen' by the tree mining algorithm. The 'external data' is used to verify the hypothesis so that it can become reliable enough to extend the current knowledge. In many cases, the choice of various data mining parameters can affect the nature and granularity of the obtained results. Where the hypothesis is not supported, the parameters will be adjusted and the previous steps alternated.

5. Illustration of the approach

A synthetic dataset was created analogous to a common representation of patient's records from the mental health domain. The XML document created consisted of 30 transactions (i.e. patient records). The document contains records for three different illness types: schizophrenia, depression, bipolar. The IMB3-Miner [8] algorithm was applied on the dataset using transactional support definition. The minimum support threshold chosen was 7. A number of subtree patterns were detected as frequent. We focused on the largest subtree patterns (i.e. containing most nodes), as the smaller subtree patterns were subsets of the larger ones. The illness type was indicated, and this allows one to associate the causal patterns with the specific illness type.

In Figure 5, we show three different patterns each associated with a specific mental illness type. As can be seen the patterns are meaningful in the sense that the causes are associated with an illness. Note that this is only an illustrative example used to clarify the principle behind the data mining application, and by no mean indicates the real causes. The real-world data would be of similar nature but much more complex in regard to the information contained within the patient's records. Furthermore, the underlying tree structure in which the information is presented can vary between different applications. This would pose no limitation for the application of our algorithm since it is well scalable for large and complex data of varying structural characteristics, as was demonstrated in [8, 10].

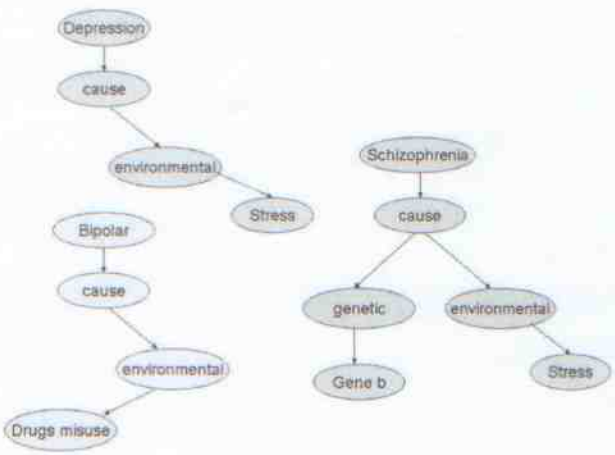

Figure 5 Data patterns derived from artificial dataset

6. Benefits and significance of the research

The number of mentally ill people is increasing globally each year. This introduces major costs in economics and human terms, to the individual communities and the nation in general, both in rural and

urban areas. The newly derived knowledge could help in prevention of mental illness and assist in delivery of effective and efficient mental health services. Various community groups could greatly benefit from information systems based on data mining technology, e.g.

- physician would receive support in early diagnosis and treatment of mental illnesses
- patients and their caregivers would receive support in dealing with, managing and treating illness;
- accurate, reliable and up-to-date information would be available for general public that will help understanding of mental illness. This will support management and prevention of mental illness.
- medical researchers would receive support in advancing their research in identifying the disease causing factors and effective patient treatments. This would reduce the possibility of redundant research (saving research time, effort and resources) and facilitate development of technologies for maintaining good health.
- the cost of the mental health budget would be significantly reduced and the spending power of money spent on health and medical research enhanced by providing better information use [30].

Such systems go some way to delivering what Patel *et al* [31] say is 'necessary to transform the quality of mental health care'. They improve the infrastructure for evidence-based interventions and provide innovation for quality improvement in mental health care.

7. Conclusion and future work

Data mining systems in general could play not only an important but a crucial role in deriving knowledge and assisting in the prevention, diagnosis, treatments and control of mental illness. In this paper, we have illustrated application of data mining technology within mental health domain. We have explained how tree mining algorithms can be effectively used for mining of patient record presented in semi-structured form.

In future, we aim to apply the developed tree mining algorithms on real world datasets and benchmark our approach with other related methods.

8. References

[1] Lopez AD, Murray CCJL (1998): The Global Burden of Disease, 1990-2020. *Nature Medicine* vol. 4, pp. 1241-1243.

[2] Holloway A, Van Laar RK, Tothill RW, and Bowtell D (2002): Options available - from start to finish – for obtaining data from DNA microarrays II. *Nature Genetics Supplement*, vol. 32, pp. 481-489.

[3] Horvitz-Lennon M, Kilbourne AM, Pincus HA (2006): From Silos To Bridges: Meeting The General Health Care Needs Of Adults With Severe Mental Illnesses. *Health Affairs* vol. 25, no. 3, pp. 659-669.

[4] Han J, Kamber M (2006): *Data Mining: Concepts and Techniques* (2nd edition). San Francisco: Morgan Kaufmann.

[5] Agrawal R, Srikant R (1994): Fast algorithms for mining association rules. *Proceedings of the International Conference of Very Large Data Bases*, Santiago, Chile, 1994.

[6] Tan H, Dillon TS, Hadzic F, Chang E (2006): SEQUEST: mining frequent subsequences using DMA Strips. *Proceedings of the International Conference on Data Mining and Information Engineering*, Czech Republic.

[7] Tan H, Dillon TS, Hadzic F, Chang E, Feng L (2005): MB3-Miner: mining eMBedded sub-TREEs using Tree Model Guided candidate generation. *Proceedings of the International Workshop on Mining Complex Data*, held in conjunction with ICDM'05, USA.

[8] Tan H, Dillon TS, Hadzic F, Feng L, Chang E (2006): IMB3-Miner: Mining Induced/Embedded subtrees by constraining the level of embedding. *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Singapore, 2006.

[9] Tan H, Dillon TS, Hadzic F, Chang E (2006): Razor: mining distance constrained embedded subtrees. *Proceedings of the IEEE ICDM 2006 Workshop on Ontology Mining and Knowledge Discovery from Semistructured documents*, China, 2006.

[10] Hadzic F, Tan H, Dillon TS (2007): UNI3 – Efficient Algorithm for Mining Unordered Induced Subtrees Using TMG Candidate Generation. *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining*, Hawaii, 2007.

[11] Hadzic F, Dillon TS, Sidhu A, Chang E, Tan H (2006): Mining Substructures in Protein Data, *Proceedings of the IEEE ICDM 2006 Workshop on Data Mining in Bioinformatics*, China.

[12] Sidhu AS, Dillon TS, Sidhu BS, Setiawan H (2004): A Unified Representation of Protein Structure Databases. *Biotechnological Approaches for Sustainable Development*, pp. 396-408.

[13] Smith DG, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR (2005): Genetic epidemiology and public health: hope, hype, and future prospects. *The Lancet*, vol. 366, no. 9495, pp. 1484-1498.

[14] Liu J, Juo SH, Dewan A, Grunn A, Tong X, Brito M, Park N, Loth JE, Kanyas K, Lerer B, Endicott J,

Penchaszadeh G, Knowles JA, Ott J, Gilliam TC, Baron M (2003): Evidence for a putative bipolar disorder locus on 2p13-16 and other potential loci on 4q31, 7q34, 8q13, 9q31, 10q21-24, 13q32, 14q21 and 17q11-12. *Mol Psychiatry*, vol. 8, no. 3, pp. 333-342.

[15] Craddock N, Jones I (2001): Molecular genetics of bipolar disorder. *The British Journal of Psychiatry*, vol. 178, no. 41, pp. 128-133.

[16] Zaki MJ, Wang JTL, Toivonen HTT (2003): Data Mining in Bioinformatics: report on BIOKDD'03'. *SIGKDD Explorations*, vol. 5, no. 2, pp. 198-200.

[17] Piatetsky-Shapiro G, Tamayo P (2003): Microarray Data Mining: Facing the Challenges. *SIGKDD Explorations*, vol. 5, no. 2, pp. 1-6.

[18] Onkamo P, Toivonen H (2006): A survey of data mining methods for linkage disequilibrium mapping. *Human genomics*, vol. 2, no. 5, pp. 336-340.

[19] Xiao Y, Yao J-F, Li Z, Dunham MH (2003): Efficient data mining for maximal frequent subtrees. *Proceedings of the IEEE International Conference on Data Mining*, USA, pp. 379-386.

[20] Asai T, Arimura H, Uno T, Nakano S (2003): Discovering Frequent Substructures in Large Unordered Trees. *Proceedings of the International Conference on Discovery Science*, Japan.

[21] Nijssen S, Kok JN (2003): Efficient discovery of frequent unordered trees. *Proceedings of the International Workshop on Mining Graphs, Trees, and Sequences*, Croatia.

[22] Chi Y, Yang Y, Muntz RR (2004): HybridTreeMiner: An efficient algorihtm for mining frequent rooted trees and free trees using canonical forms. *Proceedings of the International Conference on Scientific and Statistical Database Management*, Greece.

[23] Hido S,Kawano H (2005): AMIOT: Induced Ordered Tree Mining in Tree-structured Databases. *Proceedings of the IEEE International Conference on Data Mining*, USA, pp 170-177.

[24] Zaki MJ (2005): Efficiently Mining Frequent Trees in a Forest: Algorithms and Applications. *IEEE Transaction on Knowledge and Data Engineering*, vol. 7, no. 8, pp. 1021-1035.

[25] Zaki MJ (2005): Efficiently Mining Frequent Embedded Unordered Trees. *Fundamenta Informaticae* vol. 65, pp. 1-20.

[26] Termier A, Rousset M-C, Sebag M (2002): Treefinder: A First Step Towards XML Data Mining. *Proceedings of the International Conference on Data Mining*, Japan.

[27] Shasha D, Wang JTL, Zhang S (2004): Unordered Tree Mining with Applications to Phylogeny. *Proceedings of the International Conference on Data Engineering*, USA.

[28] Wang JTL, Shan H, Shasha D, Piel WH (2003): Treerank: A similarity measure for nearest neighbor searching in phylogenetic databases. *Proceedings of the International Conference on Scientific and Statistical Database Management*, USA.

[29] Bayardo RJ, Agrawal R, Gunopulos D (1999): Constraint-based rule mining on large, dense data sets. Proceedings of the International Conference on Data Engineering, Australia.

[30] Garber AM (2006): PERSPECTIVE: To Use Technology Better. *Health* Affairs, pp. W51-W53.

[31] Patel KK, Butler B, Wells KB (2006): What Is Necessary To Transform The Quality Of Mental Health Care. *Health Affairs* vol. 25, no. 3, pp. 681-693.