International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:1, 2008

Data Mining in Oral Medicine Using Decision Trees

Fahad Shahbaz Khan, Rao Muhammad Anwer, Olof Torgersson, and Göran Falkman

***Abstract*—**Data mining has been used very frequently to extract hidden information from large databases. This paper suggests the use of decision trees for continuously extracting the clinical reasoning in the form of medical expert's actions that is inherent in large number of EMRs (Electronic Medical records). In this way the extracted data could be used to teach students of oral medicine a number of orderly processes for dealing with patients who represent with different problems within the practice context over time.

***Keywords*—**Data mining, Oral Medicine, Decision Trees, WEKA.

I. Introduction

DATA mining has recently become very popular due to the emergence of vast quantities of data. In this paper, potential pitfalls and practical issues about data mining in oral medicine are discussed. Theoretical education in oral medicine to dental students is usually given through lectures, books and scientific papers. Text books often present a small number of cases for each diagnosis. Students may therefore receive information that does not reflect the reality a clinician in oral medicine encounters in daily practice. The learning that comes with experience from treatment outcomes may therefore be missing when the student graduates. mEduWeb is a program that was written and designed to give students the possibility to study oral medicine through a web interface [1].

mEduWebII used the Medview database which contains data from several thousand patient examinations [1]. The purpose of our work has been to seek improvements in the current mEduWebII program or, to be more specific, improvement of step-wise exercises in mEduWebII. Step-wise exercises present an orderly process for dealing with a patient who represents with a problem. The problem with step-wise exercises is that the students learn with one predefined structured thinking process for solving one type of problem. This paper identifies whether decision trees could be used for continuously extracting clinical reasoning in the form of medical expert's action that is inherent in large number of EMRs. In this way, the student would be taught a number of orderly processes for dealing with patients who represent with different types of problems. Several results have been gathered through a series of experiments.

Fahad Shahbaz Khan and Rao Muhammad Anwer are with Department of Applied IT, IT University of Göteborg, Chalmers University of Technology, Göteborg, Sweden (e-mail: fahadji@yahoo.com, raocool35@yahoo.com).

Olof Torgersson is with Department of Computer Science and Engineering, Chalmers University of Technology, Göteborg, Sweden (e-mail: oloft@cs.chalmers.se).

Göran Falkman is with School of Humanities and Informatics, University of Skövde, Skövde, Sweden (e-mail: goran.falkman@his.se).

II. Decision Trees

Decision trees are often used in classification and prediction. It is simple yet a powerful way of knowledge representation. The models produced by decision trees are represented in the form of tree structure. A leaf node indicates the class of the examples. The instances are classified by sorting them down the tree from the root node to some leaf node.

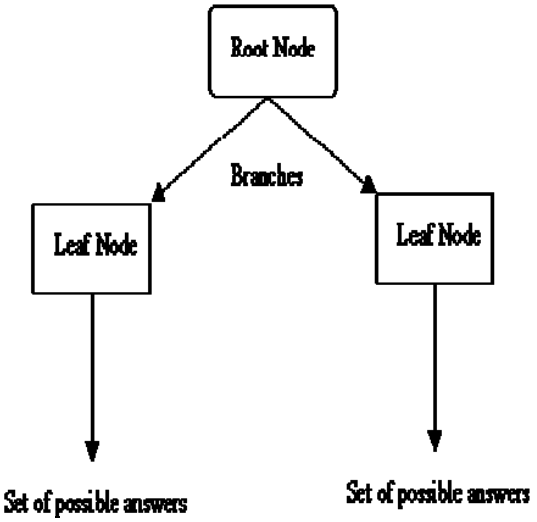

Fig. 1 A Decision Tree [2, 3, and 4]

III. Experiments and Results

We have used Weka [5] for our experiments. Weka is a collection of machine learning algorithms for data mining tasks. Weka's native storage method is ARFF format. So a conversion has been performed to make the examination data available for analysis through Weka. The most important part in the entire data mining process is preparing the input for data mining investigation. The Medview database contains data from more than 20000 patient's examinations. The data contains a lot of missing values. Graphical Visualizations in Weka make it easy to understand the data. Fig. 2 at the end of this paper (in screenshots section) shows the visualization of

some attributes from Medview database through Weka. The database contains both numeric and nominal attributes. Numeric attributes measure is either integer valued or real valued numbers. Nominal attributes take on values from a finite set of possibilities.

Decision trees represent a supervised approach to classification. Weka uses the J48 algorithm, which is Weka's implementation of C4.5 [7] Decision tree algorithm. J48 is actually a slight improved to and the latest version of C4.5. It was the last public version of this family of algorithms before the commercial implementation C5.0 was released. Originally the Medview database has data for over 180 different attributes. The significant problem has been the missing values. In Fig. 3 (in screenshots section), attribute "ADV-DRUG" is shown to have 64% missing values.

The reason for selecting C4.5 decision tree algorithm is the algorithm's ability to handle data with missing values. It also avoids overfitting the data and reduce error pruning. Initially all 180 attributes have been tested to review different results, but they could not produce the desired results. Fig. 4 (in screenshots section) shows the results of running C4.5 Decision tree algorithm.

The output shown in 4 (in screenshots section) needs some explanation to see how the tree structure is represented. Each line represents a node in the tree. The lines those that starts with a '|', are child nodes of the first line. A node with one or more '|' character before the rule is the child node of the node the right most line of '|' character terminates at. If the rule is followed by a colon and a class designation then that designation becomes the classification of the rule. If it isn't followed by a colon, continue to the next node in the tree [6].

The first series of experiments has generated faulty classification models. As a next step only those examinations have been considered that have values for the attributes "Diag-Def" and "Vis-cause= Primärundersökning". The value of Vis-cause, "Primärundersökning", corresponds to primary visits and the Diag-Def attribute corresponds to definitive diagnosis. These two attributes are known to be significant and should therefore play vital roles in the classification. Further, all those attribute have been ignored that have more than 80% missing values. Fig. 5 (in screenshots section) shows one of the results that have been generated by applying C4.5 decision tree algorithm on refined dataset.

Here the results have been somewhat similar to most of the experiments carried out earlier in the sense that those attributes which are not considered useful in diagnosis have been dominant in the decision tree model. The tree model only has one attribute and that is "P-code" which is patient identifier. This is not an important question to be asked in practice for diagnostic purpose.

The results obtained in the previous experiments have been still faulty so in the next step the advice has been taken from the domain expert. This will also prompt to follow the footsteps of the experts and how they handle a particular situation. The set of attributes have been reduced and only those have been considered that are asked in common practice. The attributes are:

- **Adv-drug**
- **Alcohol**
- **Allergy**
- **Bleed**
- **Care-provider**
- **Careprovider-now**
- **Civ-stat**
- **Diag-def**
- **Diag-hist**
- **Diag-tent**
- **Dis-now**
- **Dis-past**
- **Drug**
- **Family**
- **Health**
- **Lesn-on**
- **Lesn-site**
- **Lesn-trigg**
- **Mucos-attr**
- **Mucos-colr**
- **Mucos-site**
- **Mucos-size**
- **Mucos-txtur**
- **Ref-cause**
- **Smoke**
- **Snuff**
- **Symp-now**
- **Symp-on**
- **Symp-site**
- **Symp-trigg**
- **Treat-drug**
- **Treat-eval-obj**
- **Treat-eval-subj**
- **Vas-now**
- **Vis-cause**

As before, only those examinations have been considered which have no missing values for "Diag_def" attribute and the value of "Vis-cause = Primärundersökning". Fig. 6 (in screenshots section) shows the tree model obtained after applying the algorithm on the newly transformed dataset. In Fig. 6, "Ref-cause" is at the root of the tree and it gives information about why a certain patient has been referred to, follow by "Mucos-txtur" and so on. The derived tree structure is important in the sense that the sequence of attributes in the tree reflects the questions normally asked in practice (i.e. asking about "Mucos-txtur" gives much more information than to ask about some other attributes). The result has been much more accurate from the previous ones in the sense that the derived tree structure reflects the relative importance of examination questions asked in practice. Fig. 7 shows a small tree structure taken from the previous decision tree model reflecting the importance of questions.

Applying C4.5 to Examination Terms

Ref-cause = “Slemhinneförändring”
Mucos-txtur = “Epiteldeskvamation”: **Morsicatio K131**
Ref-cause = “Slemhinneförändring”
Mucos-txtur = “Plaque”
Smoke = “3cigaretter utan filter/dag”: **Leukoplaki homogen K132**
Ref-cause = “Slemhinneförändring”
Mucos-txtur = “Normal”
Adv-drug = “Nej”
Symp-now = “Nej”: **Frisk slemhinna K000**
Ref-cause = “Slemhinneförändring”
Mucos-txtur = “Svullnad”
Civ-stat =”Gift”: **Gingivit-plackinducerad K051**

Fig. 7 Example tree structure reflecting importance of questions asked in practice

IV. Related Work

Medview [1] was designed earlier to support the learning process in oral medicine and oral pathology. The purpose of Medview was to provide a computerized teaching aid in oral medicine and oral pathology. In this regard, a clinical database was created from the referrals and has a large variation of clinical cases displayed by images and test based information. The students reach the database through the media. They can practice and learn at any convenient time. Medview contains search tools to explore the database and the students can study single cases or analyze various clinical parameters [1]. mEduWeb [1] is a web-based educational tool that allows students to search in the database and generate exercises with pictures of real patients [1]. mEduWebII was intended to enhance and improve mEduWeb program better. It uses the MedView database containing several thousand patient examinations [1]. Our work explored the possibilities of using Data mining technique (Decision trees) on the Medview database. In this regard, a series of experiments have been performed. This can really help students in learning a number of orderly processes for dealing with patients. The final model reflects the relative importance of examination questions normally asked in practice. This will also provide the basis of evaluating the performance of students.

V. Conclusion

Initially the experiments have been conducted on the whole Medview dataset. Graphical Visualizations have been performed in order to make it easier to understand the data itself. The reason for selecting the C4.5 decision tree algorithm is because the algorithm has the ability to handle data with missing attribute values better than ID3 decision tree algorithm. It also avoids overfitting the data and reduces error pruning. The experiments involved more than 8000 examinations with 182 attributes. Each attribute has been tested to review different results but they could not produce the desired results due to a large amount of missing values in the data.

In the next step, only those examinations have been considered that have values for attributes “Diag-def” and “Vis-cause = Primärundersökning”. The value of Viscause, “Primärundersökning”, corresponds to primary visits. These two attributes are significant and plays a vital role in classification. The results have been somewhat similar to most of the experiments carried out earlier in the sense that those attributes which are not considered useful in diagnosis have been dominant in the decision tree model (i.e. in one of the experiments, the tree model only has one attribute and that is “P-code”, Patient Identifier, which is not an important question to be asked in practice for diagnostic purpose).

In the next step the advice has been taken from the domain expert. The set of attributes have been reduced and only those haven been considered which are asked in common practice. There have been improvements in the decision tree models carried out from the set of attributes given by the domain expert. Also ignoring all those examinations where the value of “Diag-def” has been missing has made a positive impact on the outcomes later on. The improved step-wise exercise presents information in the same order given by the decision tree. Figure 6 (in screenshots section) shows some part of a decision tree model. “Ref-cause” is at the root of the tree and it gives information about why a certain patient has been referred to. The model reflects the relative importance of examination questions asked in practice, e.g. to ask about “Ref-cause” and “Mucos-txtur” gives more information than to ask about “Civ-stat”. It also describes the level of difficulty in terms of relative complexity of different paths leading to terminal. This is useful to set different level of difficulties to solve a particular problem and forms the basis of evaluating the performance of students.

Acknowledgment

We would like to thank everyone involved in WEKA.

References

[1] *A Computerised Teaching Aid in Oral Medicine and Oral Pathology*. Mats Jontell, Oral medicine, Sahlgrenska Academy, Göteborg University. Olof Torgersson, department of Computing Science, Chalmers University of Technology, Göteborg.
[2] T. Mitchell, *"Decision Tree Learning"*, in T. Mitchell, Machine Learning, the McGraw-Hill Companies, Inc., 1997, pp. 52-78.
[3] P. Winston, "*Learning by Building Identification Trees*", in P. Winston, Artificial Intelligence, Addison-Wesley Publishing Company, 1992, pp. 423-442.
[4] Howard J. Hamilton’s CS Course: Knowledge Discovery in Databases. Accessed 06/06/12.
[5] http://www.cs.waikato.ac.nz/ml/weka/, accessed 06/05/21.
[6] http://grb.mnsu.edu/grbts/doc/manual/J48_Decision_T rees.html, accessed 06/06/12.
[7] Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kauffman, 1993.

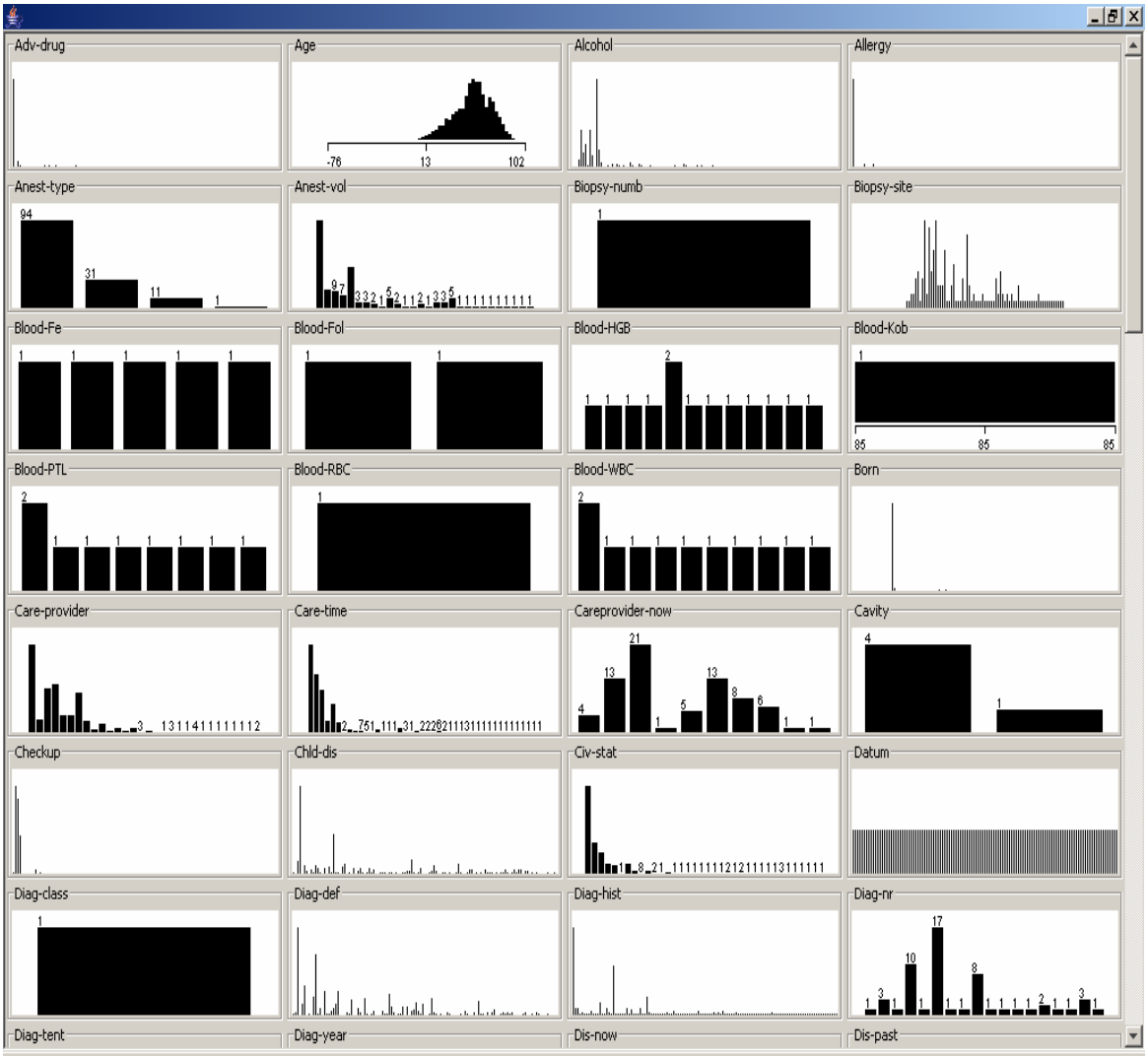

Fig. 2 Visualization of Some Attributes from medview Database through Weka

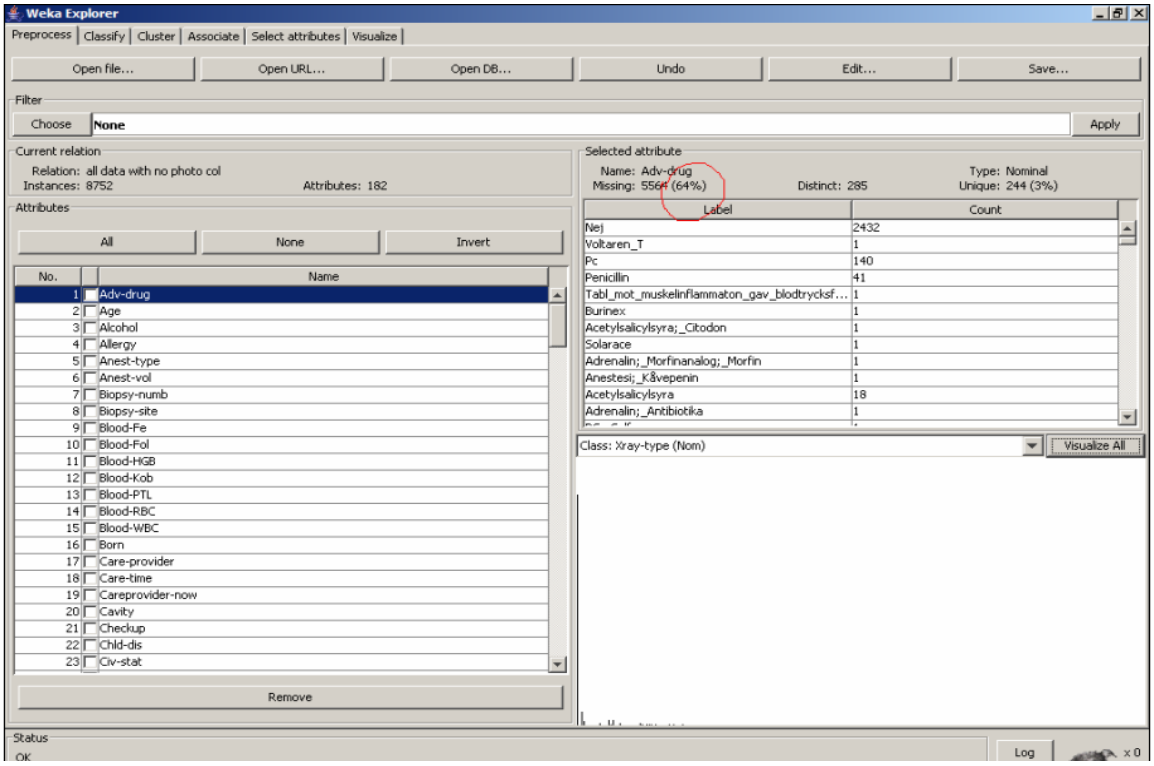

Fig. 3 Missing Values in the Attribute “ADV-DRUG”

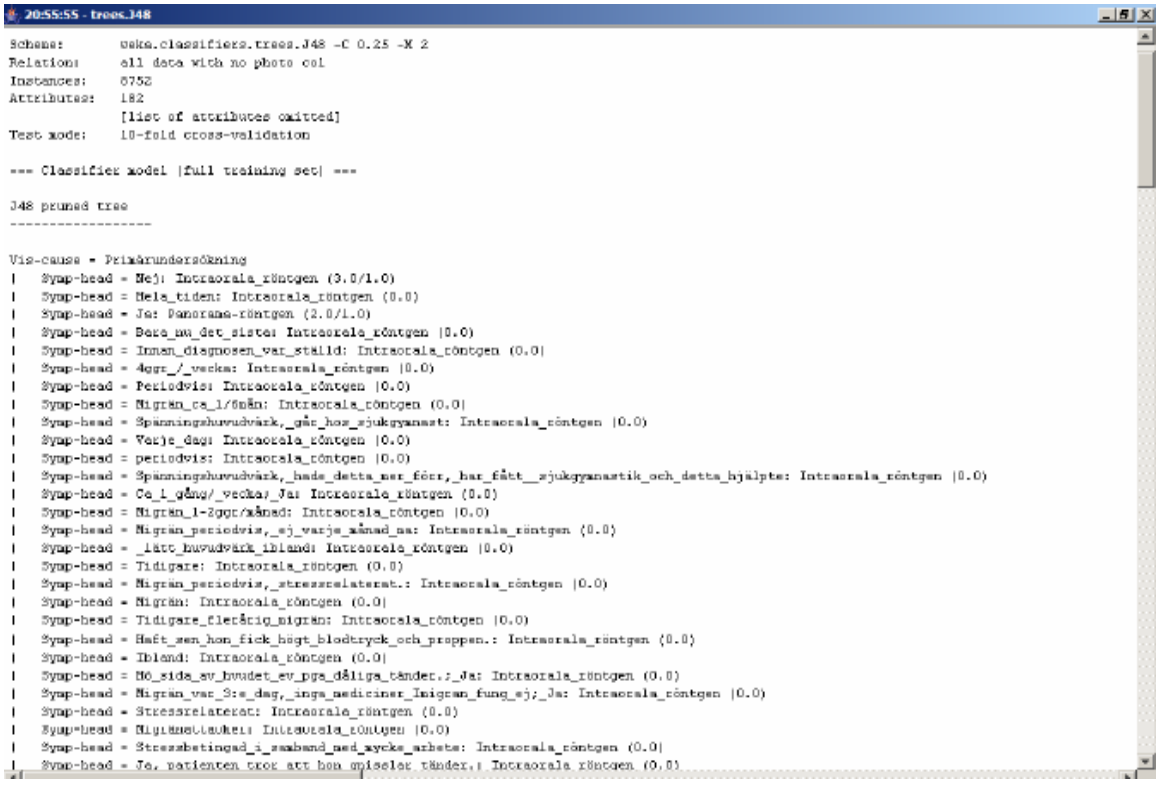

Fig. 4 Running C4.5 Decision Tree Algorithm on Examination Term

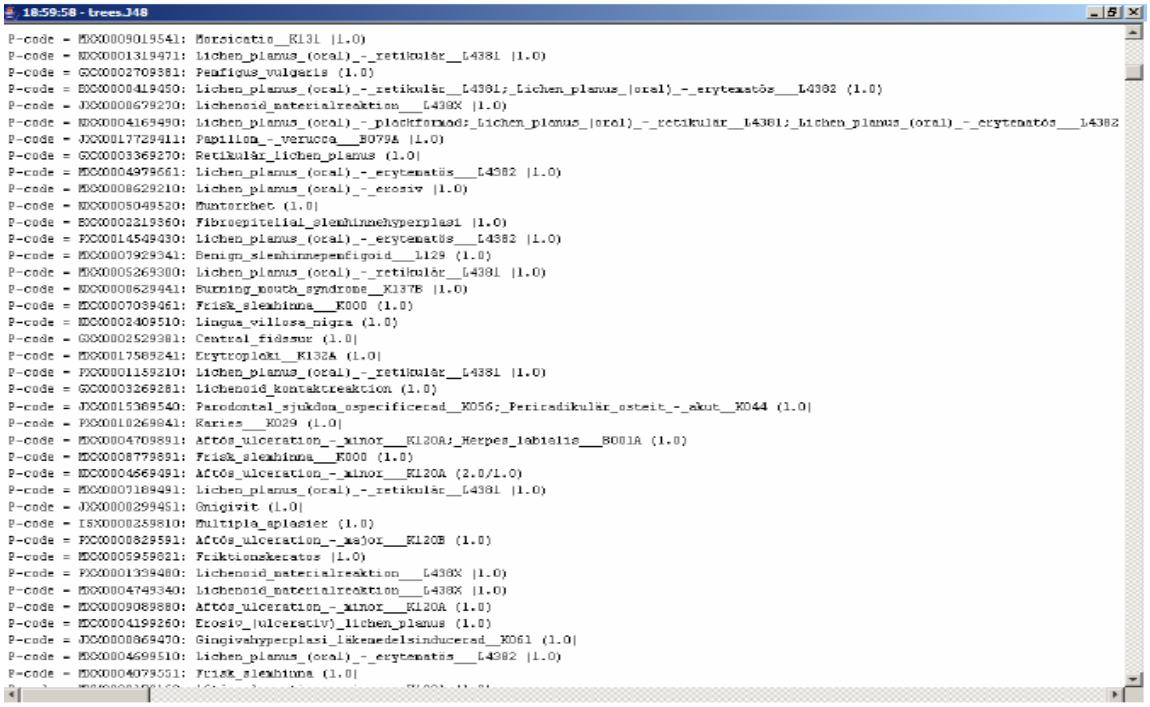

Fig. 5 Decision Tree Model Obtained on Refined Dataset

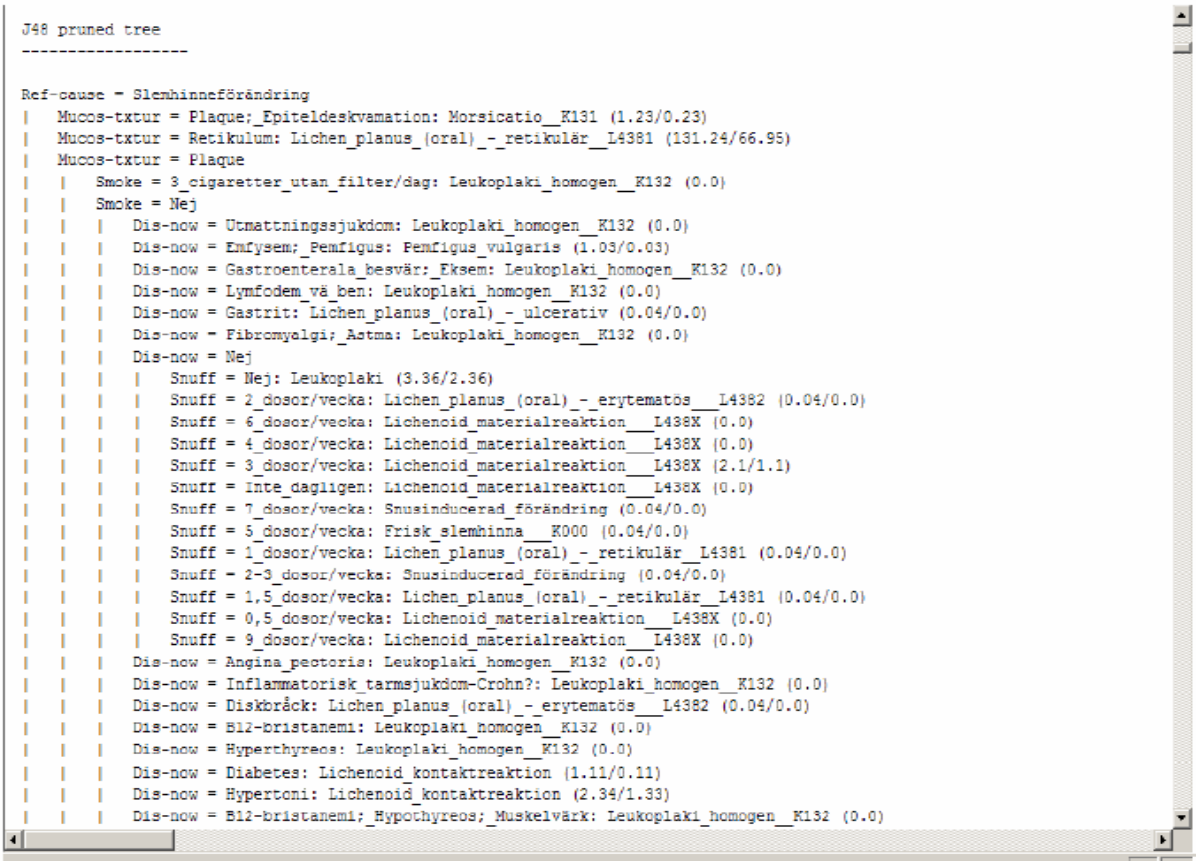

Fig. 6 The Final Decision Tree Model