Rule Extraction in Diagnosis of Vertebral Column Disease

Yaser Issam Hamodi
Department of Computer Engineering
Ministry Of Higher Education & Scientific Research
Baghdad , Iraq
yaserissam@yahoo.com

Abstract **-** Computer aided diagnosis systems are getting importance in recent years. In this paper I worked on vertebral column disease (Disc Hernia, Spondylolisthesis) biomedical Dataset for disease diagnosis. I have results with a performance comparison among some of the best data mining techniques (classifying and clustering). Different type of classification algorithms gives very promising results (more than %90 accuracy) in disease diagnosis.

Keywords *-Vertebral column disease, data mining algorithms, classification, clustering.*

*****

I. Introduction

The vertebral column is a system composed by a group of vertebras, invertebrate discs, nerves, muscles, medulla and joints [2].

As any human system spine can be exposed to the accidents which caused many medical problems, two of most common problems are: first, disc hernia occurs when the one of the inter-vertebral discs slips out from its place and compresses the nerves, so caused a lot of pain Fig. 1, And Fig. 2. shows the second disease which called Spondylolisthesis which is occurs when one vertebra slips forward over the vertebra below it. Most often, that happens in the low back (lumbar spine) because that part of the spine bears a lot of weight and absorbs a lot of directional pressures [8].

In this experimental study we have applied some of the top 10 data mining algorithms of famous dataset.

The interesting dataset is a biomedical Dataset which built by Dr. Henrique da Mota in the Group of Applied Research in Orthopaedics (GARO) of the Centre Médico-Chirurgical de Réadaptation des Massues, Lyon, France. The data have been organized in two different classes but related classification tasks. The first task consists of the classifying patients as belonging to one out of three categories: Normal (100 patients), Disc Hernia (60 patients) and Spondylolisthesis (150 patients) this data named “column_3C_weka”. For the second task, the categories Disc Hernia and Spondylolisthesis are merged into a single category labeled as 'abnormal 'this named “column_2C_weka”. Thus, the second task consists of classifying patients as belonging to one out of two categories: Normal (100 patients) or Abnormal (210 patients) and the name of our Dataset is” Vertebral Column Dataset” [11].

This Dataset contains six biomechanical attributes for each patient, those attribute are: pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius and grade of spondylolisthesis.

I have used the following labels for the classes: DH (Disc Hernia), Spondylolisthesis (SL), Normal (NO) and Abnormal (AB).

Dataset containing values for six biomechanical features used to classify patients into 3 classes (normal, disc hernia and spondilolysthesis) and 2 classes (normal or abnormal).

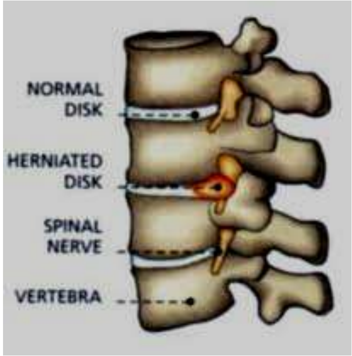

Fig. 1 Disc hernia

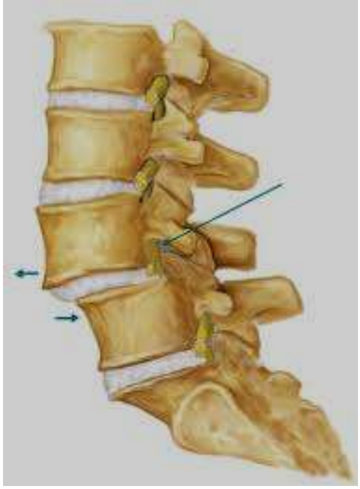

Fig. 2 Spondylolisthesis

II. Methodology

In this section I explaining the algorithms which have used in our experiment both classifications, clustering algorithms.

A. *Classification*

-Decision trees in data mining, is a predictive model which can be used to represent both classifiers and regression models. In operations research, on the other hand, decision trees refer to a hierarchical model of decisions and their consequences [5].

In other words, it is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The topmost node in a tree is the root node [4].

It is considered one of the most important, powerful and popular algorithms in knowledge analysis and data mining, it has became very effective tool in many fields such as information extraction, text recognition and machine learning.

-Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered "naïve." Bayesian belief networks are graphical models, which unlike naïve Bayesian classifiers allow the representation of dependencies among subsets of attributes [4].

-Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes.

Each tuple represents a point in an n-dimensional space. In this way, all of the training tuples are stored in an n-dimensional pattern space. When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k "nearest neighbors" of the unknown tuple. "Closeness" is defined in terms of a distance metric, such as Euclidean distance.

The Euclidean distance between two points or tuples is (1), say, X1 = (x11, x12, : : : , x1n) and X2 = (x21, x22, : : : , x2n), then

$$dist(X1, X2) = \sqrt{\sum_{i=0}^{n}(x1i - x2i)^2}. \qquad (1)$$

In other words, for each numeric attribute, we take the difference between the corresponding values of that attribute in tuple X1 and in tuple X2, square this difference, and accumulate it. The square root is taken of the total accumulated distance count [4].

$$v' = \frac{v - \min A}{\max A - \min A}. \qquad (2)$$

Equation (2) is used to prevent attributes with initially large ranges from outweighing attributes with initially smaller ranges, for example, to transform a value of a numeric attribute A to v' in the range.

B. Clustering

-Filtered Clusterer Class for running an arbitrary clusterer on data that has been passed through an arbitrary filter. Like the clusterer, the structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure.

-Expectation–maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables [10]. The algorithm is described as follows:

1. Make an initial guess of the parameter vector: This involves randomly selecting k objects to represent the cluster means or centers (as in k-means partitioning), as well as making guesses for the additional parameters
2. Iteratively refine the parameters (or clusters) based on the following two steps:
   (a) Expectation Step: Assign each object xi to cluster Ck with the probability

$$p(x_i \in C_k) = p(C_k \mid x_i) = \frac{p(C_k)p(xi \mid C_k)}{p(x_i)}, \qquad (3)$$

where p(x_i | C_k =N(m_k , E_k (x_i)) follows the normal i.e. ,caussian) distribution around mean, m_k ,with expectation, E_k . In other words, this step calculates the probability of cluster membership of object x_i , for each of the clusters. These probabilities are the "expected" cluster memberships for object x_i .

   (b) Maximization Step: Use the probability estimates from above to re-estimate (or refine) the model parameters. For example [4],

$$(m_k) = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i P(x_i \in C_k)}{\Sigma_j P(x_i \in C_j)}. \qquad (4)$$

II. Experiments and Results

Using Classification Algorithms

Applying Decision tree algorithm to the Dataset it is taken around 0.2 seconds to build model for dataset named column_2C_weka and 0.3 seconds for column_3C_weka,Table 1 shows the results we have got and follow it the rules which we could extract them from the charts .

TABLE 1. Result with Decision Trees

	column_2C_weka	column_3C_weka
Total Number of Instances	310	310
Correctly Classified Instances	91.2903 %	92.032%
Incorrectly Classified Instances	8.7097 %	7.0968 %
Number of Leaves	10	12
Size of the tree	19	23

The rules belonging to "colum_2C_weka" are:

1. IF the degree of spondylolisthesis >19.854759 THEN the case will be Abnormal (145.0/2.0)

2. IF the degree of spondylolisthesis <=19.854759 AND the pelvic radius >125.212716 THEN the case will be Normal (52.0/7.0)
3. IF the degree of spondylolisthesis <=19.854759 AND the pelvic radius <=125.212716 AND sacral slope >40.475232 AND the degree of spondylolisthesis >9.064582 THEN the case will be Abnormal (6.0/1.0)
4. IF the degree of spondylolisthesis <=19.854759 AND the pelvic radius <=125.212716 AND sacral slope >40.475232 AND the degree of spondylolisthesis <= 9.064582 AND the pelvic tilt <=18.898407 THEN the case will be Normal (2.0)
5. IF the degree of spondylolisthesis <=19.854759 AND the pelvic radius <=125.212716 AND sacral slope >40.475232 AND the degree of spondylolisthesis <= 9.064582 AND the pelvic tilt >18.898407 AND lumbar lordosis angle >56.3 THEN the case will be Normal (5.0)
6. IF the degree of spondylolisthesis <=19.854759 AND the pelvic radius <=125.212716 AND sacral slope >40.475232 AND the degree of spondylolisthesis <= 9.064582 AND the pelvic tilt >18.898407 AND lumbar lordosis angle <= 56.3 AND the pelvic incidence >6.5013773THEN the case will be Abnormal (3.0)
7. IF the degree of spondylolisthesis <=19.854759 AND the pelvic radius <=125.212716 AND sacral slope >40.475232 AND the degree of spondylolisthesis <= 9.064582 AND the pelvic tilt >18.898407 AND lumbar lordosis angle <= 56.3 AND the pelvic incidence <= 6.5013773THEN the case will be Normal (3.0/1.0)
8. IF the degree of spondylolisthesis <=19.854759 AND the pelvic radius <=125.212716 AND sacral slope <= 40.475232 AND the pelvic tilt >9.976664 THEN the case will be Abnormal (62.0/16.0)
9. IF the degree of spondylolisthesis <=19.854759 AND the pelvic radius <=125.212716 AND sacral slope <= 40.475232 AND the pelvic tilt >9.976664 AND the pelvic radius >115.877017 THEN the case will be Normal (9.0)
10. IF the degree of spondylolisthesis <=19.854759 AND the pelvic radius <=125.212716 AND sacral slope <= 40.475232 AND the pelvic tilt >9.976664 AND the pelvic radius <= 115.877017 THEN the case will be Abnormal (5.0)

And the rules belonging to “colum_3C_weka” are :

1. IF the degree of spondylolisthesis >15.779697 THEN the case will be Spondylolisthesis (148.0/3.0)
2. IF the degree of spondylolisthesis <=15.779697 AND sacral slope >46.636577 AND the degree of spondylolisthesis >8.235294 THEN the case will be Spondylolisthesis (4.0/1.0)
3. IF the degree of spondylolisthesis <=15.779697 AND sacral slope >46.636577 AND the degree of spondylolisthesis <= 8.235294 THEN the case will be Normal (21.0/1.0)
4. IF the degree of spondylolisthesis <=15.779697 AND sacral slope <= 46.636577 AND the pelvic radius <= 117.422259 THEN the case will be Hernia (46.0/12.0)
5. IF the degree of spondylolisthesis <=15.779697 AND sacral slope <= 46.636577 AND the pelvic radius > 117.422259 AND sacral slope >28.131342 AND pelvic tilt <=12.306951 THEN the case will be Normal (33.0)
6. IF the degree of spondylolisthesis <=15.779697 AND sacral slope <= 46.636577 AND the pelvic radius > 117.422259 AND sacral slope >28.131342 AND pelvic tilt > 12.306951 AND the degree of spondylolisthesis <= 5.074353 THEN the case will be Normal (24.0/3.0)
7. IF the degree of spondylolisthesis <=15.779697 AND sacral slope <= 46.636577 AND the pelvic radius > 117.422259 AND sacral slope >28.131342 AND pelvic tilt > 12.306951 AND the degree of spondylolisthesis > 5.074353 AND the degree of spondylolisthesis <= 8.235294 THEN the case will be Hernia (6.0)
8. IF the degree of spondylolisthesis <=15.779697 AND sacral slope <= 46.636577 AND the pelvic radius > 117.422259 AND sacral slope >28.131342 AND pelvic tilt > 12.306951 AND the degree of spondylolisthesis > 5.074353 AND the degree of spondylolisthesis > 8.235294 THEN the case will be Normal (5.0/1.0)
9. IF the degree of spondylolisthesis <=15.779697 AND sacral slope <= 46.636577 AND the pelvic radius > 117.422259 AND sacral slope <= 28.131342 AND pelvic tilt > 17.114312 THEN the case will be Hernia (10.0)
10. IF the degree of spondylolisthesis <=15.779697 AND sacral slope <= 46.636577 AND the pelvic radius > 117.422259 AND sacral slope <=28.131342 AND pelvic tilt <= 17.114312 AND the pelvic tilt >14.930725 THEN the case will be Normal (4.0)
11. IF the degree of spondylolisthesis <=15.779697 AND sacral slope <= 46.636577 AND the pelvic radius > 117.422259 AND sacral slope <=28.131342 AND pelvic tilt <= 17.114312
12. AND the pelvic tilt <= 14.930725 AND the degree of spondylolisthesis >0.75702 THEN the case will be Hernia (5.0)
13. IF the degree of spondylolisthesis <=15.779697 AND sacral slope <= 46.636577 AND the pelvic radius > 117.422259 AND sacral slope <=28.131342 AND pelvic tilt <= 17.114312

AND the pelvic tilt <= 14.930725 AND the degree of spondylolisthesis <= 0.75702 THEN the case will be Normal (4.0/1.0)

Applying Naïve Bayes to the Dataset it have taken 0.02 seconds to get the results for dataset named column_2C_weka and 0.03 seconds for column_3C_weka,Table 2 shows the results :

TABLE 2. Results with Naïve Bayes Classifier

	column_2C_weka	column_3C_weka
Total Number of Instances	310	310
Correctly Classified Instances	78.0645%	83.5484 %
Incorrectly Classified Instances	21.9355 %	16.4516 %

Table 3 shows the results after applying IBk algorithm (k-nearest neighbor's classifier), The time taken to build model is 0.02 seconds for column_2C_weka and 0.03 seconds for the other one.

TABLE 3. Result with k-nearest neighbor classifier

	column_2C_weka	column_3C_weka
Total Number of Instances	310	310
Correctly Classified Instances	100 %	100 %
Incorrectly Classified Instances	0 %	0 %

Using Clustering Algorithms

Second part of our experiment is using Data mining clustering algorithms, Fig. 3 shows the result after applying FilteredClusterer to column_2C_weka which has taken time to build model (full training data) 0.07 seconds, while Fig.4 shows the result for column_3C_weka after having undergone the same algorithm where has taken 0.08 seconds to build model.

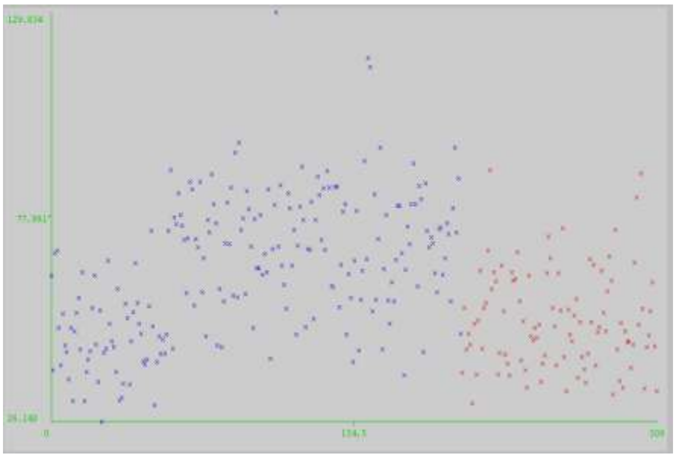

Fig. 3 Applying FilteredClusterer on column_2C_weka result

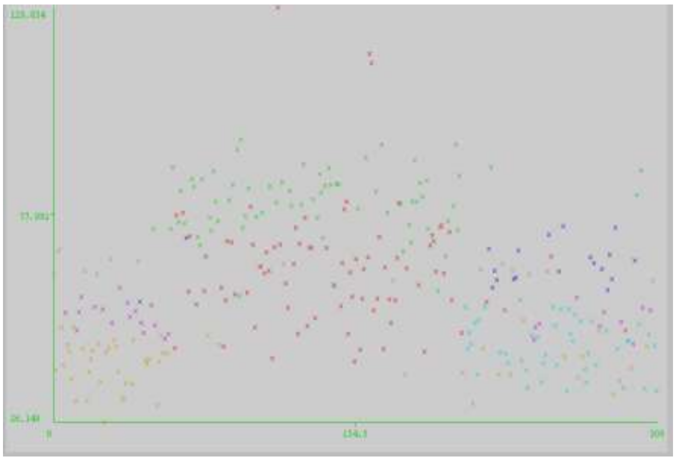

Fig. 4 Applying FilteredClusterer on column_3C_weka result

From table 4 we can note the number of the instances in each cluster.

TABLE 4. Result with Filtered Cluster Algorithm

	Clustered Instances	
	1	**2**
column_2C_weka	210	100
column_3C_weka	159	151

EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters this algorithm is the second clustering algorithm we have used in our experiment, Fig.5 and Fig.6 shows the results. note that applying EM has registered -23.03206 as LOG Likelihood and it has taken 59.88 seconds with column_2C_weka and 43.21 seconds for column_3C_weka with LOG Likelihood=-23.16856.

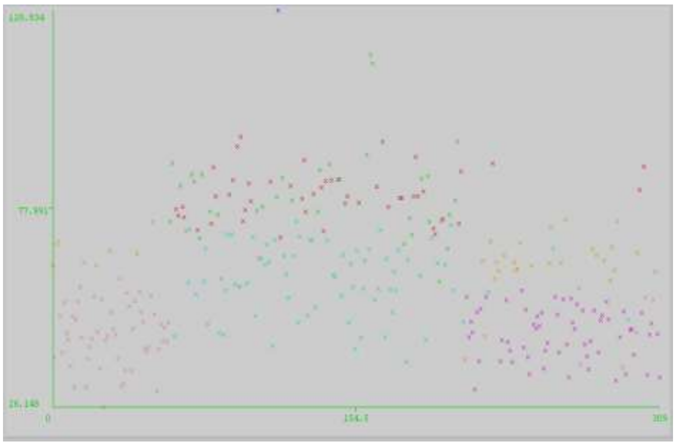

Fig. 5 Applying EM algorithm on column_2C_weka result

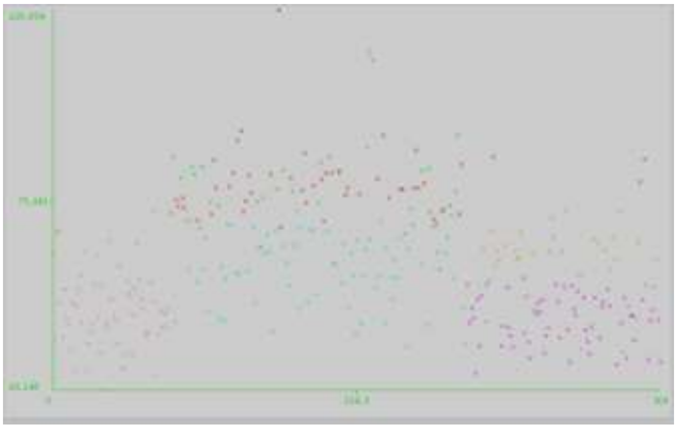

Fig. 6 Applying EM algorithm on column_3C_weka result

In the following table (Table 5) appears the number of instances in each cluster.

TABLE 5. Result with EM Algorithm

	Clustered Instances	
	column_2C_weka	**column_3C_weka**
0	20	1
1	77	46
2	70	34
3	45	70
4	16	59
5	28	62
6	51	35
7	3	

SUMMERY

Today early medical diagnosis is very important. From other disciplinary researchers also they work on this issue to help doctors in their decision. Our dataset has been subject to many of experiments and all of them aim to evaluate the interest of embedded options methodologies for aiding the diagnostic of pathology on the Vertebral Column. Data mining based algorithms have been given very promising result for classification. In the same goal this paper with its contained results may consider a basis for helping researchers and developers to apply further experiments.

REFERENCES

[1] A. R. Rocha Netu e G. A. Barreto, Member IEEE,"On the Application of Ensembles of Classifiers to the Diagnosis of Pathologies of the Vertebral Column: A Comparative Analysis," *IEEE LATIN AMRICA TRANSACTIONS*, vol. 7, no.4, AUG 2009.

[2] Ajalmar R. da Rocha Neto (1), Ricardo Sousa(2), Guilherme de A. Barreto(1), and Jaime S. Cardoso(2), "Diagnostic of Pathology on the Vertebral Column with Embedded Reject Option," (1)Depto. Engenharia de Teleinform_atica, Universidade Federal do Cear_a (UFC), (2) INESC Porto, Faculdade de Engenharia da Universidade do Porto, Portugal.

[3] Berthonnaud, E., Dimnet, J., Roussouly, P., Labelle, H.:"Analysis of the sagittal balance of the spine and pelvis using shape and orientation parameters. Journal of Spinal Disorders & Techniques 18(1)," 40-47 (2005).

[4] Jiawei Han and Micheline Kamber,"Data Mining Concepts Techniques,"2nd ed.

[5] Lior Rokach (1) ,Oded Maimon (2)," DATA MINING WITH DECISION TREES Theory and Applications,"(1) Ben-Gurion University, Israel,(2) Tel-Aviv University, Israel, Series in Machine Perception and Artificial Intelligence - Vol. 69.

[6] XindongWu , Vipin Kumar , J. Ross Quinlan , Joydeep Ghosh , Qiang Yang , Hiroshi Motoda , Geoffrey J. McLachlan , Angus Ng , Bing Liu , Philip S. Yu , Zhi-Hua Zhou , Michael Steinbach , David J. Hand and Dan Steinberg," Top 10 algorithms in data mining," SURVEY PAPER, Knowl Inf Syst (2008) 14:1–37 DOI 10.1007/s10115-007-0114-2, Springer-Verlag London Limited 2007, Published online: 4 December 2007.

[7] Matthew J. Beal," VARIATIONAL ALGORITHMS FOR APPROXIMATE BAYESIAN INFERENCE," *The Gatsby Computational Neuroscience Unit University College London*, M.A., M.Sci., Physics, University of Cambridge, UK (1998).

[8] Kamiah A. Walker; Reviewed by Jason M. Highsmith, MD," What Is Spondylolisthesis? ".

[9] Alexis Roche,"EM algorithm and variants: an informal tutorial." Service Hospitalier Fr´ed´eric Joliot, CEA, F-91401 Orsay, France,Spring 2003.

[10] Maya R. Gupta and Yihua Chen, "Theory and Use of the EM Algorithm." Foundations and Trends, In Signal Processing, vol. 4, No. 3 (2010) 223–296

[11] http://archive.ics.uci.edu/ml/datasets/Vertebral+Column# .