JHI JOURNAL OF HEALTH INFORMATICS

Association and Classification Data Mining Algorithms Comparison over Medical Datasets

Comparação de Algoritmos de Classificação e de Associação de Mineração de Dados sobre Bases de Dados Médicos

Comparación de Algoritmos de Clasificación y de Asociación de Míneria de Datos sobre Bases de Datos Médicos

Bruno Fernandes Chimieski[1], **Rubem Dutra Ribeiro Fagundes**[2]

Descritores: Data Mining; Classification; Association

ABSTRACT

Objectives: Compare Data Mining algorithms related to Classification and Association tasks over medical datasets about dermatology, vertebral column and breast cancer patients, analyzing which is the best one over each of these datasets. **Methods:** The classification algorithms are ran over these datasets and compared using precision, F-measure, ROC curve and Kappa performance metrics. For associaton task, the Apriori algorithm is ran to get a significant number of rules with confidence above 90%. **Results:** For diagnostics prediction about breast cancer and dermatology issues, the best classification algorithm was BayesNet and for vertebral column was the Logistic Model Tree. For association task, were extracted 100 knowledge rules for breast cancer and dermatology issues with confidence higher than 90% while for vertebral column were found 18 with same confidence. **Conclusion:** The comparison was useful to prove the possibility of using Data Mining algorithms to help Medicine decision engine with good precision.

Keywords: Mineração de Dados; Classificação; Associação

RESUMO

Objetivos: Compar os algoritmos de Mineração de Dados de Classificação e Associação de dados sobre bases de dados de dermatologia, câncer de mâma e de problemas da coluna vertebral. **Métodos:** Os algoritmos de classificação foram executados sobre essas bases de dados e comparadas pelas métricas de precisão, *F-measure,* curva ROC e *Kappa.* Para associação, o algoritmo *Apriori* é executado para gerar um número significante de regras com confiança acima de 90%. **Resultados:** Para a predição de diagnósticos sobre câncer de mâma e dermatologia o melhor algoritmo foi o BayesNet e para coluna vertebral foi o de Árvore de Modelo Logístico. Para a tarefa de associação, foram extraídas 100 regras de conhecimento para a base de câncer de mâma e de dermatologia com confiança acima de 90% enquanto para a da coluna vertebral foram encontradas 18 com a mesma confiança. **Conclusão:** A comparação foi útil para provar a possibilidade do uso de algoritmos de Mineração de Dados no auxílio ao processo decisório na Medicina com boa precisão.

Descriptores: Minería de Datos; Clasificación; Asociación

RESUMEN

Objetivos: Comparar los algoritmos de minería de datos relacionados con las tareas de clasificación y asociación de conjuntos de datos médicos sobre dermatología, coluna vertebral y patientes con cáncer de mama, analizando cual es el mejor en cada uno de estos conjuntos de datos. **Métodos:** Los algoritmos de clasificación se pasó por encima de estos conjuntos de datos y se compararon con las métricas de rendimiento precisión, F-medida, la curva ROC y Kappa. Para la tarea Associaton, el algoritmo Apriori obtiene normas de confianza superior al 90%. **Resultados:** Para la predicción de diagnóstico sobre el cáncer de mama y problemas dermatológicos el mejor algoritmo de clasificación fue BayesNet y de la columna vertebral era el árbol del modelo logístico. Para tarea de asociación, se extrajeron 100 reglas de conocimiento para el cáncer de mama y problemas dermatológicos con confianza mayor que 90%, mientras que para la columna vertebral se encontraron 18 con la misma confianza. **Conclusión:** La comparación es útil para demostrar la posibilidad de utilizar algoritmos de minería de datos para ayudar a motor de decisóin de Medicina con buena precisión.

[1] *Graduate student (MSc) in Electrical Engineering from the Graduate Program in Mathematical Sciences and Technology, Catholic University of Rio Grande do Sul - PUCRS, Porto Alegre (RS)*
[2] *Ph.D. in Electrical Engineering from the University of São Paulo - USP, São Paulo (SP), Brazil.*

Autor Coorespondente: **Bruno Fernandes Chimieski**
e-mail: **bruno.chimieski@gmail.com**

Artigo recebido: **09/10/2012**
Aprovado: **05/03/2013**

INTRODUCTION

Information Technology applied to Medical Environments

The amount of data acquired electronically from patients has grown exponentially during the past decades. Hospital databases including demographic systems, electronic patient records, as well as order-entry, laboratory, pharmacy, and radiology systems grow in scope and data capacity with each year. In its raw form, data are relatively uninformative. Handled properly, however, data can be mined for novel and unexpected information(1). Decision support tools have been used in prevention, drug prescription, diagnosis and disease management. Adherence to preventive practices in areas such as vaccinations, screening, and cardiovascular risk reduction has been improved when decision support tools have been integrated into patient care. Also, Computer Science has developed a number of new tools to extract information from data and enhance analysis by the human clinical expert(1). More recently, machine learning techniques, such as neural networks, have been used to detect myocardial infarctions, breast cancer, cervical cancer, and nosocomial disease outbreaks. Additionally, decision support tools can be kept current by people or technologies dedicated to that task, and clinicians can tap in on that current information on demand, rather than needing to maintain their own, memory-based version of current medical knowledge. There are limited data as to the benefit of DSSs (Decision Support Systems), but studies have shown that computerized systems can improve clinician performance and positively affect patient outcomes. There is also evidence that DSSs can improve the eficiency of care by reducing the amount of time clinicians spend on administrative tasks and the turn-around time between test ordering and performance. DSSs have also been shown to reduce the costs of medical care. The cognitive component of a DSS can use empirical knowledge about the association between diseases and symptoms and this knowledge(1). DSS must begin with a knowledge base, use some kind of an "engine" and produce or effect recommendations or interventions. The "engine" is the underlying software and analysis methodology(1). The KDD (Knowledge Discovery over Databases) process involves using the database along with any required selection, preprocessing, subsampling, and transformations of it; applying Data Mining methods (algorithms) to enumerate patterns from it; and evaluating the products of Data Mining to identify the subset of the enumerated patterns deemed knowledge. The Data Mining component of the KDD process is concerned with the algorithmic means by which patterns are extracted and enumerated from data(2). In this context, this research tries to compare the classification and association algorithms that are used by Weka Data Mining workbench in order to evaluate which of them are the best ones to deliver the right diagnostics of selected diseases, and to generate a set of knowledge rules with high confidence to help doctors to learn more from their patients.

METHODS

Data Mining and Weka

As the world grows in complexity, overwhelming us with the data it generates, Data Mining becomes our only hope for elucidating the patterns that underlie it. Intelligently analyzed data is a valuable resource. It can lead to new insights and, in commercial settings, to competitive advantages. Data Mining is about solving problems by analyzing data already present in databases. There are different styles of learning appear in Data Mining applications. For this paper, are considered two of them, which are the classification and association learning. The classification learning is the learning scheme presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples. Classification learning is sometimes called supervised because, in a sense, the method operates under supervision by being provided with the actual outcome for each of the training examples. In association learning, any association among features is sought, not just ones that predict a particular class value. Association rules differ from classification rules in two ways: they can "predict" any attribute, not just the class, and they can predict more than one attribute's value at a time. Because of this there are far more association rules than classification rules. For this reason, association rules are often limited to those that apply to a certain minimum number of examples—say 80% of the dataset—and have greater than a certain minimum accuracy level— say 95% accurate. Even then, there are usually lots of them, and they have to be examined manually to determine whether they are meaningful or not. Association rules usually involve only nonnumeric attributes(3).

WEKA

Experience shows that no single machine learning scheme is appropriate to all Data Mining problems. The universal learner is an idealistic fantasy, because real datasets vary, and to obtain accurate models the bias of the learning algorithm must match the structure of the domain. So, Data Mining is an experimental science. The Weka workbench is a collection of state-of-the-art machine learning algorithms and data preprocessing tools. It is designed so that you can quickly try out existing methods on new datasets in flexible ways. It provides extensive support for the whole process of experimental Data Mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning. As well as a wide variety of learning algorithms, it includes a wide range of preprocessing tools. This diverse and comprehensive toolkit is accessed through a common interface so that its users can compare different methods and identify those that are most appropriate for the problem at hand. Weka was developed at the University of Waikato in New Zealand, and the name stands for Waikato Environment for Knowledge Analysis(3). The version of Weka used in this research was 3.6.7.

Overview of Mined Datasets

All of the following datasets were accessed from the Machine Learning Repositoy of University of California Irvine, and are all of public domain for research purposes.

Breast Cancer

This breast cancer database was obtained from the University of Wisconsin Hospitals, in Madison, Winsconsin, USA from Dr. William H. Wolberg and was donated in 15 July 1992 for UCI Machine Learning Repository. Each instance has one of 2 possible classes: benign or malignant tumor. The total number of instances are 699 and the number of attributes is 10 plus the class attribute. The attributes are explained in the following table:

Table 1 - Attributes of Breast Cancer Dataset1 Attributes of Breast Cancer Dataset

Attribute	Domain
Sample code number	Id number
Clump Trickness	1-10
Uniformity of Cell Size	1-10
Uniformity of Cell Shape	1-10
Marginal Adhesion	1-10
Single Epithelial Cell Size	1-10
Bare Nuclei	1-10
Bland Chromatin	1-10
Normal Nucleoli	1-10
Mitoses	1-10
Class	(2 for benign, 4 for malignant)

Dermatology

The original owners of this dataset are Nilsel Ilter, M.D., Ph.D., from Gazi University, School of Medicine and H. Altay Guvenir, PhD., from Bilkent University, Department of Computer Engineering and Information Science, all from Ankara, Turkey. The dataset was donated in January, 1998. This database contains 34 attributes, 33 of which are linear valued and one of them is nominal, and 366 instances. The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. The list of attributes is the following: erythema, scaling, definite, itching, koebner, polygonal, follicular, oral, knee, scalp, family, melanin, eosinophils, PNL, fibrosis, exocytosis, acanthosis, hyperkeratosis, parakeratosis, clubbing, elongation, thinning, spongiform, munro, focal, disappearance, vacuolisation, spongiosis, saw-tooth, horn, perifollicular, inflammatory, band-like and age. The class distribution is as in the table below:

Table 2 - Attributes of Dermatology Dataset2 Attributes of Dermatology Dataset

Class code	Class	Number of instances
1	Psoriasis	112
2	Seboreic dermatitis	61
3	Lichen planus	72
4	Pityriasis rósea	49
5	Cronic dermatitis	52
6	Pityriasis rubra pilaris	20

Vertebral Column

The donors of this dataset are Guilherme de Alencar Barreto and Ajalmar Rêgo da Rocha Neto from the Department of Teleinformatics Engineering, of the Federal University of Ceará, Fortaleza, Ceará, Brazil, and Henrique Antonio Fonseca da Mota Filho from the Hospital Monte Klinikum, Fortaleza, Ceará, Brazil. In this paper, was used the version of the dataset with 3 categories, Normal (100 patients), Disk Hernia (60 patients) or Spondylolisthesis (150 patients). Each patient is represented in the data set by six biomechanical attributes derived from the shape and orientation of the pelvis and lumbar spine (in this order): pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius and grade of spondylolisthesis.

Classification Algorithms

The algorithms that were evaluated in this paper belong to two main types of classifications models, based on decision trees and on bayesian classifiers.

A decision tree is a flow-chart-like tree structure or model of decisions, where each internal node denotes a test on an attribute, each branch represents an outcome of the test that leads to a leaf node, representing classes or class distributions. The topmost node in a tree is the root node. Decision trees are constructed in a top-down recursive divide-and-conquer manner. Starting with a training set of tuples and their associated class labels, the training set is recursively partitioned into smaller subsets as the tree is being built. Nevertheless, not all branches are seen in a decision tree. Tree pruning attempts to identify and remove branches that may reflect noise or outliers, with the goal of improving classification accuracy[4].

LMT

A logistic model tree basically consists of a standard decision tree structure with logistic regression functions at the leaves, much like a model tree is a regression tree with regression functions at the leaves. As in ordinary decision trees, a test on one of the attributes is associated with every inner node. For a nominal (enumerated) attribute with k values, the node has k child nodes, and instances are sorted down one of the k branches depending on their value of the attribute. For numeric attributes, the node has two child nodes and the test consists of comparing the attribute value to a threshold: an instance is sorted down the left branch if its value for that attribute is smaller than the threshold and sorted down the right branch otherwise[5].

Bayesian classifiers

Bayesian classifiers are statistical classifiers based on Baye's theorem that predict the probability of a tuple to belong to a certain class. Similarly to decision trees and selected neural network classifiers, when applied to large databases, Bayesian classifiers (as the Naïve Bayesian and Bayesian Networks) show high accuracy and speed. Naïve Bayesian classifiers assume class conditional independence, meaning that the effect of an attribute value on a given class is independent of the values of the other attributes[4].

Bayesian Network

The Naïve Bayes classifier produces a probability estimate, rather than hard classifications. For each class value, it estimates the probability of a given tuple to belong to that class. Futhermore, for a given class of a tuple, it assumes that the attributes are conditionally independent of each other, simplifying the computing. Developed by Pearl (1995), bayesian networks (BN), also known as Bayes nets, are a statistical based alternative belonging to the family of probabilistic graphical models that represent a set of random variables in the nodes and their conditioned dependencies in the edges between the nodes, combining principles from graphical theory, probability theory, computer sciences and statistics[4].

Evaluation criteria for classification

Most of the analysis of evaluation starts from a confusion matrix, which displays the amount of correct and incorrect classifications from each class. The true positives (TP) and true negatives (TN) are correct classifications. A false positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A false negative (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive. The true positive rate is TP divided by the total number of positives, which is TP + FN; the false positive rate is FP divided by the total number of negatives, FP + TN. The overall success rate is the number of correct classifications divided by the total number of classifications:

$$\frac{TP+TN}{TP+TN+FP+FN} \qquad (5)$$

Finally, the error rate is one minus this[3].

Accuracy

The accuracy of a classifier is the percentage of correctly classified instances in a test set, measuring how well the classifier recognizes instances of the various classes[4].

ROC

The Receiver Operating Characteristic (ROC) curve is a graphical plot of sensitivity given by true positive rate as a function of false positive rate, being a good tool for visualizing a classifier performance and to select a suitable decision threshold. The area under curve (AUC) of ROC is often used as a statistic for model comparison. The larger AUC is, the more accurate the classifier is. For example, an ideal classifier has an AUC of 1 while a poor one has an area of 0.5[4].

Kappa

The overall percent of correctly classified instances reflects a simple evaluation of a classifier, the same as evaluation by the area under a receiver operating curve (ROC). Because a classifier relying on random selection of instances will frequently classify some instances correctly, the kappa statistic is used to control for those instances that may have been correctly classified only by chance. Also, can be evaluated the accuracy of each classifier by its F-measure, which represents the harmonic mean between precision and recall[6]. Kappa statistic is used to assess the accuracy of any particular measuring cases, it is usual to distinguish between the reliability of the data collected and their validity. The average Kappa score from the selected algorithm is around 0.6-0.7[7].

F-measure

The F-Measure was used, because despite Precision and Recall being valid metrics in their own right, one can be optimised at the expense of the other. The F-Measure only produces a high result when Precision and Recall are both balanced, thus this is very significant[8]. Precision is the proportio of the predicted relevant pages that were

$$\frac{TP}{FP+TP} \qquad (6)$$

Recall, is the proportion of the relevant pages that were correctly idenified:

$$\frac{TP}{FN+TP} \qquad (7)$$

F-Measure, is derived rom precision and recall values:

$$\frac{2*Recall*Precision}{Recall+Precision} \qquad (8)$$

Association with Apriori algorithm

One of the most popular Data Mining approaches is to find frequent itemsets from a transaction dataset and derive association rules. Finding frequent itemsets (itemsets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori is a seminal algorithm for finding frequent itemsets using candidate generation . It is characterized as a level-wise complete search algorithm using anti-monotonicity of itemsets, "if an itemset is not frequent, any of its superset is never frequent". By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. Let the set of frequent itemsets of size k be F_k and their candidates be C_k . Apriori first scans the database and searches for frequent itemsets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent itemsets.

1. Generate C_k+1, candidates of frequent itemsets of size k + 1, from the frequent itemsets of size k.
2. Scan the database and calculate the support of each candidate of frequent itemsets.
3. Add those itemsets that satisfies the minimum support requirement to F_k+1.

Finally, many of the pattern finding algorithms such as decision tree, classification rules and clustering techniques that are frequently used in Data Mining have been developed in machine learning research community. Frequent pattern and association rule mining is one of the few exceptions to this tradition. The introduction of this technique boosted Data Mining research and its impact is tremendous. The algorithm is quite simple and easy to implement. Experimenting with Apriori-like algorithm is the first thing that data miners try to do[9].

Interestingness Measures for Association Rules

The algorithms of rules discovering make use of interesting measures in order to decrease the number of rules generates in the output of its algorithms. The interesting measures universally most used are support and confidence. The support is a measure that evaluates the frequency with the terms of a rule appears in data. In other words, the number of transactions in which the items present in the rule appears at the same time in the data. The confidence is a measure that refers to a correspondence value between the items that compose a rule. So, it express the percentage of transactions in which, having the antecedent, the consequent also exists[10].

Transformation of Datasets

For this work was needed to transform the datasets to a file format accepted by Weka, so making possible to preprocess and mine these data with it. So, the data was organized manually in csv files.

Filtering for Classification and Association Tasks

Before execute the step of mining the data during the KDD process, there is a task of preprocessing that, in this case, will be responsible of removing missing instances and removing atributes related to identification numbers of patients that are not useful for the mining task. Additionally, specifically for classification task, was needed to perform a conversion of numeric to nominal attributes, and for association task was applied a discretization over the attributes. All these preprocessing actions were done using the available features of Weka workbench.

Methods of Comparison Between Classifiers

By the adoption of Weka workbench, were choosed from the available classification algorithms, the ones that were capable of deal with the types of data used in the selected datasets. To choose the best classification method for each dataset individually (breast cancer, dermatology diseases and vertebral column issues) were selected 12 algorithms of classifiers based on decision trees and 3 based on bayesian models. Then, was choosed the best method among the 12 decision trees and the best among the 3 bayesian ones. Next, these 2 algorithms were compared in order to finally decide which was the best for the dataset experimented. The list of algorithms of decision trees were the following: Best-Firts Decision Trees, Decision Stump, Functional Trees, J48, J48 graft, Logistic Model Tree, Naive-Bayes Trees, Random Forest, Random Trees, SimpleCart, Logic-Boost Alternating Decision Tree and REPTree (Fast Decision Tree).

The list of bayesian classifiers are the following: BayesNet, NaiveBayes and NaiveBayesUpdateable. The criteria used for all the comparisons was the set of measures composed by percent of instances correctly classified, kappa statistic, area under ROC curve and F-measure. The test approach used was based in a split test with ratio 66/33. In other words, the training of each algorithm's model was done using 66% of the dataset while the testing of this trained model was done with the remaining 33% of the dataset.

Extraction of Association Rules

For this task was choosed the Apriori algorithm, by the reasons mentioned early in this paper. The rules generation in Apriori was based on confidence metric. So, were generated a set o 100 rules using the minimum metric of 0.9 of confidence.

RESULTS AND DISCUSSION

In this section will be reported the results of the experiments done with classification and association algorithms with the objective to choose the best of them to be applied on breast cancer, dermatology and vertebral column datasets.

Results for Classification Algorithms

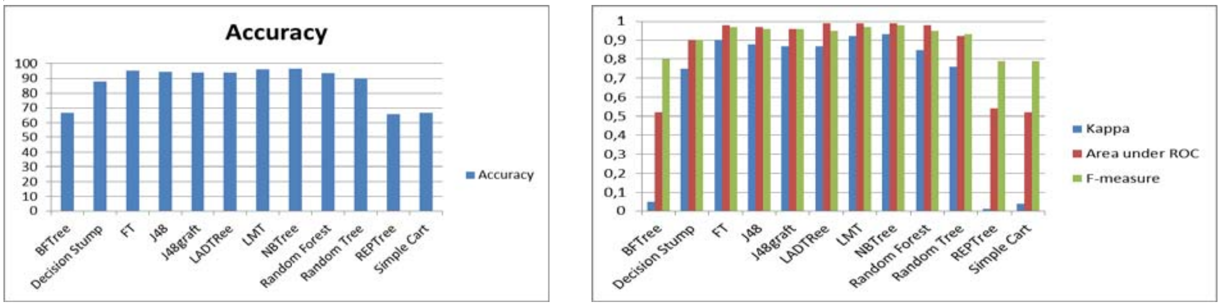

Figure 1 - Results of Decision Trees Algorithms Over Breast Cancer Dataset3 Results of Decision Trees Algorithms Over Breast Cancer Dataset

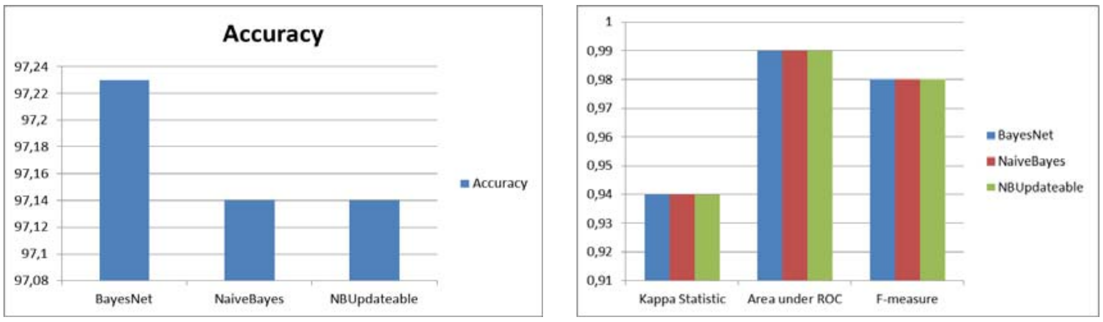

Figure 2 - Results of Bayesian Classifiers Over Breast Cancer Dataset4 Results of Bayesian Classifiers Over Breast Cancer Dataset

Discussion about Experiments for Breast Cancer Dataset

Analysing the results for breast cancer from figure 3 and 4, is noticed that the best decision tree algorithm was the NBTree, while the best bayesian classifier was BayesNet. Finally, the comparison between these algorithms reveals that BayesNet is the best one, having the same values for F-measure and area under ROC, but with higher values for kappa statistic and accuracy than NBTrees.

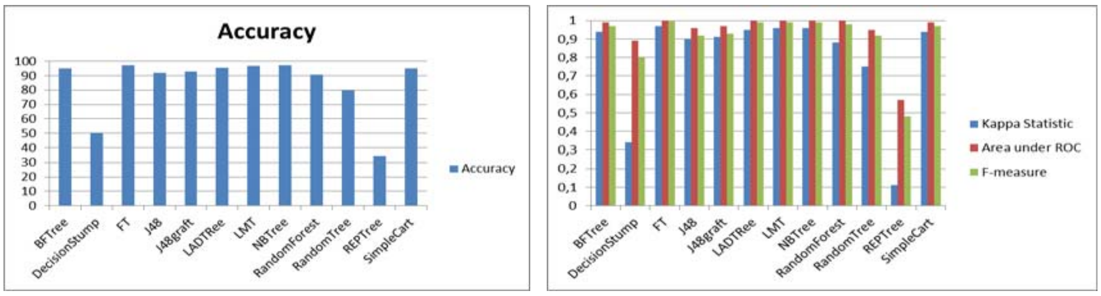

Figure 3 - Results of Decision Trees Algorithms Over Dermatology Dataset5 Results of Decision Trees Algorithms Over Dermatology Dataset

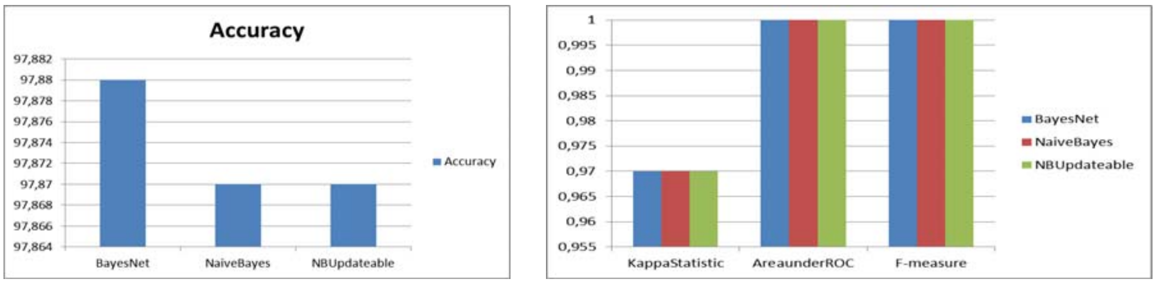

Figure 4 - Results of Bayesian Classifiers Over Dermatology Dataset6 Results of Bayesian Classifiers Over Dermatology Dataset

Discussion about Experiments for Dermatology Dataset

From the results of figure 5 and 6, the best decision tree classifier for dermatology dataset was FT. Among the bayesian classifiers, BayesNet was the better one by a difference of only 0.1 in accuracy. Comparing these two algorithms, BayesNet has advantage in accuracy by 0.5 points over FT decision tree. For other criteria, these two algorithms were equivalent. So, the best algorithm for dermatology dataset classification tasks od BayesNet.

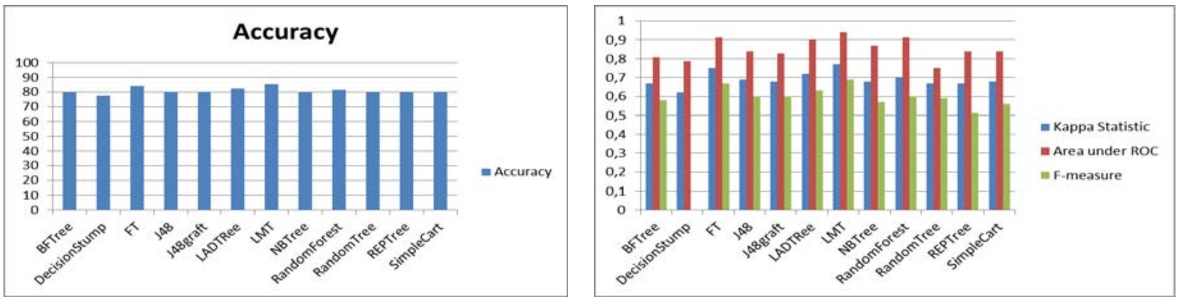

Figure 5 - Results of Accuracy for Decision Trees Algorithms Over Vertebral Column Dataset7 Results of Accuracy for Decision Trees Algorithms Over Vertebral Column Dataset

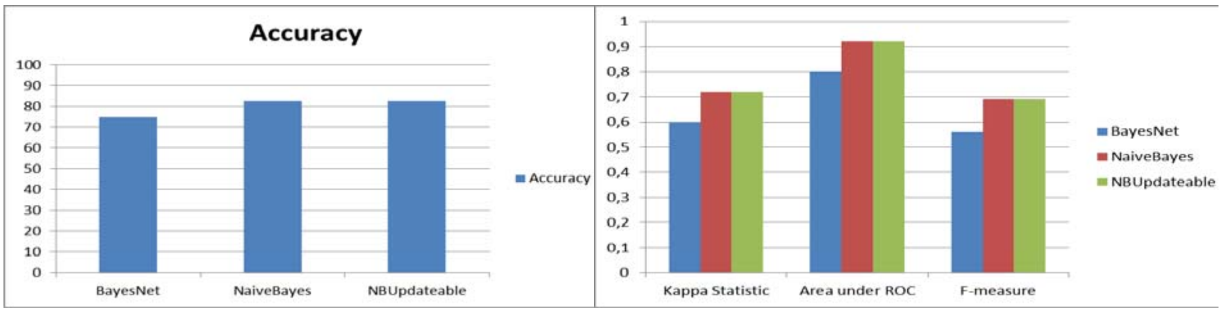

Figure 6 - Results of Accuracy for Bayesian Classifiers Over Vertebral Column Dataset8 Results of Accuracy for Bayesian Classifiers Over Vertebral Column Dataset

Discussion about Experiments for Vertebral Column Dataset

From the results of graphics 7 and 8, the best decision tree classifier for vertebral column dataset was the LMT one. Among bayesian classifiers, both NaiveBayes and NaiveBayesUpdateable had the same performance for all criteria and were better than BayesNet. Comparing these two bayesian classifiers with LMT results, was noticed that LMT had higher values for accuracy, kappa statistic and area under roc curve. Then, LMT was the best classifier for vertebral column classification task.

Results for Association Algorithms

As mentioned before, the generation of association rules was done considering as metric a minimum confidence of 0.9, and resctricting the number of rules to 100.

Here in this paper will are showed 10 rules among the 100 association rules generated, due to paper size limitations.

Association Rules Generated for Breast Cancer Dataset

For the breast cancer dataset, the Apriori algorithm obtained 100 rules with confidence higher them 0.9. The rule #100 was extracted with confidence of 0.93. So, 10 examples of the best rules found were the following ones:
1. UniformityofCellSize = '(-inf-1.9]' BareNuclei = '(-inf-1.9]' ==> Class = 2 (Conf: 1.00)
2. UniformityofCellSize = '(-inf-1.9]' BareNuclei = '(-inf-1.9]' NormalNucleoli = '(-inf-1.9]' ==> Class = 2 (Confidence: 1.00)
3. BareNuclei = '(-inf-1.9]' NormalNucleoli = '(-inf-1.9]' Mitoses = '(-inf-1.9]' ==> Class = 2 (Confidence: 0.99)
4. MarginalAdhesion = '(-inf-1.9]' BareNuclei = '(-inf-1.9]' Class = 2 ==> Mitoses = '(-inf-1.9]' (Confidence: 0.99)
5. UniformityofCellSize='(-inf-1.9]' BareNuclei='(-inf-1.9]' NormalNucleoli='(-inf-1.9]' ==> Mitoses='(-inf-1.9]' Class=2 (Confidence:0.98)

Association Rules Generated for Dermatology Dataset

For dermatology dataset, were extracted the following rules
1. follicular = '(-inf-0.3]' horn = '(-inf-0.3]' ==> perifollicular = '(-inf-0.3]' (Confidence: 1.00)
2. vacuolisation = '(-inf-0.3]' ==> melanin = '(-inf-0.3]' (Confidence: 1.00)
3. oral = '(-inf-0.3]' melanin = '(-inf-0.3]' ==> polygonal = '(-inf-0.3]' (Confidence: 1.00)
4. polygonal = '(-inf-0.3]' melanin = '(-inf-0.3]' ==> oral = '(-inf-0.3]' (Confidence: 1.00)
5. polygonal = '(-inf-0.3]' oral = '(-inf-0.3]' ==> melanin = '(-inf-0.3]' (Confidence: 1.00)

Association Rules Generated for Vertebral Column Dataset

For vertebral column dataset, was tried to obtain 100 association rules, but the confidence metric of 0.9 stopped the rules generation in 18 rules. In other words, there were only 18 possible association rules to be extracted from this dataset with confidence higher than 0.9. These 18 rules are showed below:
1. class = Normal ==> degree_spondylolisthesis = '(-inf-31.901947]' (Confidence: 1.00)
2. degree_spondylolisthesis = '(31.901947-74.862073]' ==> class = Spondylolisthesis (Confidence: 1.00)
3. class = Hernia ==> degree_spondylolisthesis = '(-inf-31.901947]' (Confidence: 1.00)
4. sacral_slope = '(34.979458-45.785721]' class = Normal ==> degree_spondylolisthesis = '(-inf-31.901947]' (Confidence: 1.00)
5. pelvic_radius = '(116.576808-125.875654]' class = Normal ==> degree_spondylolisthesis = '(-inf-31.901947]' (Confidence: 1.00)

CONCLUSION

In this paper were done experiments with Weka Machine Learning Tool in order to choose the best Data Mining algorithms to be applied over selected datasets. These algorithms will be used further to compose a Java-based Data Mining application capable of perform classification and association tasks over medical datasets about breast cancer, dermatology and vertebral column issues. This Java-based application will be developed using the Weka development API classes. Also, this application will be responsible to interact with an Android mobile application giving to it the results of diagnostics prediction (classification task) and the knowledge got from datasets in the form of a set of rules (association task). From the results achieved from this research work, for diagnostics prediction about breast cancer and dermatology issues was noticed that the best classification algorithm is BayesNet, that classified unseen instances of these datasets with accuracy higher than 97%. For vertebral column diagnostics prediction the best algorithm was the Logistic Model Tree, that classified test instances with accuracy of 85.52%. For association task, were extracted a total of 100 knowledge rules for breast cancer and dermatology issues with confidence higher than 0.9 points, or 90%. In the other hand, for the dataset about vertebral column were found 18 knowledge rules with confidence higher than 90%. About the performance of these algorithms, the time processing and CPU usage, over the selected algorithms were insignificant, due to the small number of instances and attributes of them. As Data Mining approach is a very experimental science, the result of this study related to the defintion of the best algorithm to be used cannot be extended to other databases without some experimentation over them. This happens because of the distinct nature of instances and attributes of datasets, that can lead other algorithms to perform better than the best ones related in this research. In parallel with these experiments done, was done in the first sections of this paper, an introduction about the use of decision support systems in the Medical area, showing that these systems are based in the KDD approach which has as one if its steps the Data Mining processing.

REFERÊNCIAS

1. Hanson CW. Healthcare informatics. EUA: McGraw Hill Professional; 2006 [cited 2012 Dec 17]. Available from: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Healthcare+Informatics#5
2. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI magazine [Internet]. 1996;17(3):37–54. Available from: http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1230
3. Witten I, Frank E. Data Mining: practical machine learning tools and techniques [Internet]. 2005 [cited 2012 Dec 17]. Available from: http://books.google.com/books?hl=en&lr=&id=QTnOcZJzlUoC&oi=fnd&pg=PR17&dq=Data+Mining+Practical+Machine+Learning+Tools+and+Techniques&ots=3glBgnXiOe&sig=Pq5koeF3ZT4iAlVdN-iGcVT5B48
4. Vieira JAP. Algorithm development for physiological signals analysis and cardiovascular disease diagnosis - a data mining approach [thesis]. Coimbra:University of Coimbra; 2011.
5. Landwehr N, Hall M, Frank E. Logistic model trees. Machine learning [Internet]. 2005;59(1-2):161-205. Available from: http://www.springerlink.com/index/10.1007/s10994-005-0466-3
6. Tighe P, Laduzenski S, Edwards D, Ellis N, Boezaart AP, Aygtug H. Use of machine learning theory to predict the need for femoral nerve block following ACL repair. Pain medicine (Malden, Mass.) [Internet]. 2011;12(10):1566-75. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21899717
7. Othman MF, Yau TMS. Comparison of different classification techniques using WEKA for breast cancer. 3rd Kuala Lumpur International Conference on Biomedical Engineering [Internet]. 2007;15:520–3. Available from: http://www.springerlink.com/index/G05P87776R59T378.pdf
8. Xhemali D, Stone RG. Naïve bayes vs. decision trees vs. neural networks in the classification of training web pages. *IJCSI*. 2009;4(1):16-23.
9. Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining [Internet]. Knowledge and information systems. 2007 [cited 2012 Nov 1]. p. 1–37. Available from: http://www.springerlink.com/index/10.1007/s10115-007-0114-2
10. Lucas JP. Mineração de dados apoiada pela descoberta de subgrupos através do pós-processamento de regras de associação [monografia]. Pelotas (RS): Universidade Federal de Pelotas, Instituto de Física e Matemática, Departamento de Informática; 2006.