Data mining application with case based reasoning classifier for breast cancer decision support

Abdeldjalil Khelassi
Computer Science Department
Abou Bakr Belkaied University, Temcen, Algeria
Tel: 00213550168666
Khelassi.a@gmail.com

ABSTRACT

Cytology is a complex diagnosis task which requires both expertise and experience of an oncologist for providing the cancer class and stage which is very useful in the therapy and in the surgery intervention. A case based reasoning classifier is developed with specialized agents for recognizing the malignant breast cancer. The proposed application implements a data mining method for the knowledge extraction and discovery by mining a medical database, which contains classified instances characterized by some features extracted automatically from the cytological image of the patient cancer. An original technique is implemented for enriching the retrieving process on the developed CBR system; this technique is based on the combination of global-local similarity measures and fuzzy sets for modeling the unknown response generated from the agents which increase significantly the accuracy of the system. The features selection and weighting is done by a machine learning algorithm. The efficiency of the proposed methodology has been validated through some empirical experiments applied in the cited data set which demonstrates that the developed approach achieves such average accuracies better than the current state-of-the-art approaches.

Keywords: Case Based Reasoning; data mining, Fuzzy sets; Breast cancer; Cytology.

I. INTRODUCTION

In recent years data mining has become a very popular technique for extracting information from the database in different areas due to its flexibility of working on any kind of databases and also due to the surprising results.[5] Of the data mining techniques developed recently, several major kinds of data mining methods, including generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization and meta-rule guided mining, are cited in [2]. Data classification is an important topic in the field of data mining due to its wide applications. A number of related methods have been proposed based on the well-known learning models such as decision tree or neural network [3]. Standard commercial tools such as SPSS, SAS, Clementine or even freely available tools such as WEKA implement diverse mining techniques and algorithms, some of which are very sophisticated. Given the availability and the reliability of these tools, they are preferred by practitioners and researchers over proprietary self-development tools [4]. Also some of these techniques are implemented in the sophisticated database systems as IBM DB2, Microsoft SQL Server, MYSQL, and ORACLE.[5]

The case based reasoning approach is widely and successfully applied in many domains as games, recommendation systems, information retrieval, bioinformatics, industrial applications and others. It represents a good and easy method of knowledge extraction, discovery and modelling. It consists of using the prior similar cases for resolving the newest problems.[7]

The data mining and knowledge discovery techniques are widely used to diagnose human disease. [6] In this survey a data mining application is developed by applying a case based reasoning classifier, and machine learning algorithms to determine the class of breast cancer from a pattern extracted from the image of cancerous cellular tissue see Figure1. We have also proposed a modified similarity measures, by combining the traditional similarity functions with the fuzzy sets, for ensuring a flexible and accurate model.

II. AROUND THE PROBLEM

The problem treated in this work focus on the recognition of malignant cases of breast cancer by using the extracted knowledge from a data set constructed from some sample of breast cancer cytological image. In this section we will present the needed fundament around this problem. First of
all we will present the application domain, following this, we will explain the relationship between data mining and case based reasoning, next we will present the fuzzy sets.

II. I. BREAST CANCER CAUSES AND DIAGNOSIS

The Cancer is a class of diseases in which a group of cells display uncontrolled growth (division beyond the normal limits), invasion (intrusion on and destruction of adjacent tissues), and sometimes metastasis (spread to other locations in the body via lymph or blood). These three malignant properties of cancers differentiate them from benign tumors, which are self-limited, and do not invade or metastasize. Cancer affects people at all ages with the risk for most types increasing with age [8]. Cancer caused about 13% of all human deaths in 2007[10], and more than 30% of cancer deaths can be prevented. The breast cancer represent (519 000 deaths) [9]. Risk factors include: tobacco use, being overweight or obese, low fruit and vegetable intake, physical inactivity, alcohol use, sexually transmitted HPV-infection, urban air pollution, indoor smoke from household use of solid fuels. [9]

Early detection of breast cancer is enhanced and unnecessary surgery avoided by diagnosing breast masses from Fine Needle Aspirates (FNA's)[12]. Many techniques are used in the diagnosis of breast cancer as well as Microarray [15], Cytology [12] [13][14] and Mammography [11].

The cytology or histology analysis is widely used in many hospitals for the breast cancer diagnosis and exactly for the recognizing of the malignant class. It is an indispensable exam before any surgical interventions.

Figure 1. Sample of Microscopic image of benign and malignant cancer tissue

II. II. DATA MINING WITH CASE BASED REASONING

The case based reasoning approach is widely and successfully applied in many domains as games, recommendation systems, information retrieval, bioinformatics, industrial applications and others. It represents a good and easy method of knowledge extraction, discovery and modelling.

The CBR is an intelligent approach inspired from many disciplines it draws a human reasoning model [7]. It consists of using the prior expertise to resolve a new problem. This expertise is stored as a set or collection of cases called cases base. Each case represents one problem associated with its solution. The main idea of case based reasoning is that two similar problems have the same solutions or the solution can be generated from the similar problems.

Data Mining is defined as “the process of extracting trends or patterns from data” (Wright,1998). It allows a search, for valuable information, in large volumes of data (Weiss & Indurkhya, 1998). The explosive growth in databases has created a need to develop technologies that use information and knowledge intelligently. Therefore, Data mining techniques has become an increasingly important research area(Fayyad, Djorgovski, & Weir, 1996).[2]

The case based reasoning is successfully integrated and combined with the data mining[6,17] in deferent domain of application as bioinformatics in [6]and [17], and well cited in the realized data mining applications in [2].

II. III. FUZZY SETS

The fuzzy sets [16] generalize the classical sets by considering the membership as a graded concept. The membership degree of an element x to a fuzzy set A denoted by \( \mu_A(x) \), take a value in the interval [0,1]. In this section we try to define the characteristics of membership function cited on figure2.

Definition1. The support of a fuzzy set A is the crisp set that contains all the elements of X that have nonzero membership grades in A. 
\[
\text{supp}(A) = \{x \in X, \mu_A(x) > 0\}
\]

Definition2. The core of a normal fuzzy set A is the crisp set that contains all the elements of X that have the membership grades of one in A. 
\[
\text{core}(A) = \{x \in X, \mu_A(x) = 1\}
\]

Definition3. The boundary is the crisp set that contains all the elements of X that have the
membership grades of \(0 < \mu_A(x) < 1\) in \(A\).

**Definition 4.** Having two fuzzy sets \(A\) and \(B\) based on \(X\), then both are similar if:

\[ \text{Core}(A) = \text{Core}(B) \quad \text{and} \quad \text{Supp}(A) = \text{Supp}(B) \]

**Definition 5.** If the support of a normal fuzzy set consists of a single element \(x_0\) of \(X\), which has the property \(\text{Supp}(A) = \text{Core}(A) = \{x_0\}\), then this set is called a singleton and the membership function is called triangular function.

**Definition 6.** The \(\alpha\)-cut of a fuzzy set \(A\) is:

\[ \alpha A = \{x \mid \mu_A(x) \geq \alpha\} \]

And a strong \(\alpha\)-cut

\[ \alpha^+ A = \{x \mid \mu_A(x) > \alpha\} \]

**Definition 7.** The uncertainty of a fuzzy set \(A\) is measured with the Hartley function \(U\):

\[
U(P(A)) = \frac{1}{h(A)} \left[ \log_2 |\alpha A| \right]
\]

Where \(|\alpha A|\) denotes the cardinality of the \(\alpha\)-cut \(h(A)\) the height of \(A\). This function is called also the non-specificity function.

**III. PROBLEM FORMULATION**

In order to obtain more objective and precise measurements, Dr. William H. Wolberg in University of Wisconsin Hospital, have constructed a dataset. With an image analysis program, known as Xcyt\(^1\), analyzed in [18,19,20], some cellular features is computed from 699 images of malignant and benign cancerous tissues. The data set pattern contains 11 features: The id number, the class (Benign, Malignant) and 9 attribute which describes the morphology of the cancerous cells. This data set contains 699 instances in which there are 65.5% benign and 34.5% malignant. It contains also 16 instances that contains a single missing (i.e., unavailable) attribute value, denoted by “?”. The software is available for execution over the Internet, providing previously unavailable predictive accuracy to remote medical facilities. The new images will be characterized with this tool in order to generate the same pattern of data.

The problem focused and analyzed in this work consists of extracting knowledge from the cited datasets by using a self-developed cases based reasoning classifier for generating an accurate decision and a clear explanation.

**IV. THE PROPOSED SOLUTION**

**IV.I. THE CLASSIFIER**

As shown on the figure 3, the developed classifier contains two kinds of agents Adaptation agent and Similarity agent. Each case base contains cases from the same class. Each agent uses a predefined knowledge which contains ontology, rules and heuristics to realize their local goals. The main goal of the system is to generate the class of the query but each agent is autonomic for satisfying its local goals. Each agent has a specific General User Interface for introducing the classification parameters. The interfaces also is very useful for the explanation step and for retaining new cases.

![Figure 3: The conceptual model of the classification system. SKB, AKB: Similarity, Adaptation Knowledge Base. GUI: General User Interface.](http://dollar.biz.uiowa.edu/xcyt/)

**IV.II. FUZZY SIMILARITY MEASURES**

A modified similarity measures is developed, for enriching the retrieving process in the similarity agents, by combining the traditional similarity functions with the fuzzy sets. The novel similarity function consists of defining three fuzzy sets similar \(S\), not similar \(N\) and unknown \(U\) with the following membership functions \(\mu_S, \mu_N\) and \(\mu_U\)

\[
\mu_S(x) = \begin{cases} 
0 & \text{if } x \leq a \\
\frac{x-a}{a-b} & \text{if } x > a 
\end{cases}
\]

\[
\mu_N(x) = \begin{cases} 
0 & \text{if } x \geq b \\
\frac{x-b}{b-a} & \text{if } x < b 
\end{cases}
\]

\[
\mu_U(x) = \begin{cases} 
\frac{x-b}{a-x} & \text{if } x > b \text{ ou } x < a \\
\frac{x-a}{b-x} & \text{if } x < a \text{ and } x > 0.5 
\end{cases}
\]

We have used triangular functions for representing the fuzzy sets. The variable \(x\)
represents the result of the similarity function between the query and the case by using the selected function (sigmoid, exponential, linear or the threshold). The support of the fuzzy sets is defined intuitively by using the agents GUI. The sigmoid Similarity function is defined as:

\[ \text{Sim}(Q, C) = \sum_{i=1}^{N} w_i \frac{1}{1 + e^{-\alpha (c_i - q_i)}} \]  

where \( N \) the number of attributes, \( w_i \) the weight of the feature \( A_i \). The parameters \( \alpha \) and \( \theta \) are defined intuitively after some experiments. The user can also select the distance function Euclidian or logarithmic function.

The logarithmic distance function

\[ \sigma(q_i, c_i) = \begin{cases} \ln(c_i) - \ln(q_i) & \text{for } q_i, c_i > 0 \\ \ln(-c_i) - \ln(-q_i) & \text{for } q_i, c_i < 0 \\ \text{Undefined} & \text{else} \end{cases} \]  

The Euclidian distance function

\[ \sigma(q_i, c_i) = |q_i - c_i| \]  

The uncertainty of the similarity measures is computed by the formula (1), each similarity have an appropriate uncertainty function: \( U(A_g)=U(S) \) where \( S \) represent the similar fuzzy set.

**IV. III. FEATURES SELECTION AND WEIGHTING**

The features selection plays a crucial role in the pattern modeling and classification. First it can simplify the model and this way computational cost can be reduced and also when the model is taken for practical use fewer inputs are needed which means in practice that fewer measurements from new samples are needed. Second by removing insignificant features from the data set one can also make the model more transparent and more comprehensible, providing better explanation of suggested diagnosis, which is an important requirement in medical applications. Feature selection process can also reduce noise and this way enhances the classification accuracy. In this project we have used a machine learning algorithms which implement the gradient descent approach with the following criteria:

- Performance function:
  Let \( q_i, c_1 \) and \( c_2 \) from the same class.
  \[ P(q, c_1, c_2) : \text{D}_q \times \text{D}_c \times \text{D}_c \rightarrow \{0, 1\} \]
  \[ P(q, c_1, c_2) = 1 - \text{sim}(q, c_1) - \text{sim}(q, c_2) \]  
- The learning rate=0.1.

After applying this algorithm with 100 instances from each class for the definition of the selected features and the importance degree which represent the features weighting, we have obtained the following weights (in the figure4).

**IV. IV. CASE BASES LEARNING AND OPTIMIZATION**

After defining the features weights for each similarity agents we have tried to define the case bases. First of all we have randomly stored some cases in the case base, after that we have increased the number of these cases by respecting the symmetry in the case bases i.e the same case in both case bases and we have measured the rate of correct classification for each experiment with the same test instances. The obtained results are described in the following bar chart Figure.5.
and proves that the size of the training base is not significant in some cases.

We have used also another method for defining the cases bases by applying the case base learning algorithms CBL1 and CBL2 defined in [19].

![Figure 6. Number of cases in the cases base after using the case bases Learning algorithms CBL1 and CBL2.](image)

The Case base learning algorithm CBL2 which consists to store just the non-similar cases in the case base has optimized the case base to 43% of the data base without changing the precision of the classification. We can infer from this that the CBL2 optimization algorithm increases the performance of our system by 57%.

In these experiments we have used all instances of the data sets for the learning step and 200 cases was tested. We have obtained the same rate of correct classification 100% for both experiments, this prove that the used optimization algorithm CBL2 is very useful and powerful.

V. RELATED WORKS

V.I. BREAST CANCER DIAGNOSIS

The used data set has been tested by O. L. Mangasarian and W. H. Wolberg[18][19][20] by using the linear programming approach with just 50% of dataset for the learning and the testing steps and they have obtained 93.5% as a rate of correct classification. The second use of this database is in [20] also by the same researchers by applying the nearest neighbor algorithm with just 50% of the constructed data they have obtained 93.5% as a rate of correct classification.

With our developed method by applying different machine learning algorithms we have obtained different rate of correct classification from 98% to 100%. The obtained results prove that our classifier with the proposed fuzzy similarity measures is very robust and better than the state of the art algorithms.

V.II. FUZZY CASE BASED REASONING

The case based reasoning approach has a wide use in many domains; it includes also a variety of implementations and models. Many works in the CBR combine with the traditional CBR some intelligent approaches and algorithms for resolving some specific problems or for enriching the CBR model. We can cite here some works which combine between the fuzzy approach and the Case based reasoning as [21] in which they incorporate the traditional case base paradigm by the Fuzzy Logic concepts in a flexible, extensible component-based architecture. Also [23] which enforce the case based reasoning by a fuzzy logic system. We cite also [22] in which they introduce a fuzzy model for the representation of a CBR system. Another work [24] introduces in the traditional process of CBR a Data fuzzification stage for more flexible and accurate models. In our approach we combine the fuzzy sets and the global-local similarity measures for generating three responses Similar, not similar and unknown for more transparency, flexibility and accuracy.

VI. CONCLUSION

This paper presents a successful data mining application by applying a case based reasoning for the classification of a breast cancer pattern. A machine learning algorithm was used by mining from an international database for knowledge extraction and discovery. The obtained results justify the impact of the developed method for improving the precision and the transparency of the decision support.

REFERENCES


15. Isabel Bechinarditz,Ammalia Anest “Case based reasoning with Bayesian Model Averaging: An improved method for survival analysis on Microarray Data” ICCBR10, 18th Int Conf in Case Based Reasoning, Alexandria, Italy, LNAI6176 Springer (2010).


