Hybrid Reliability-Similarity-Based Approach for Supervised Machine Learning

Walid Cherif

Abstract—Data mining has, over recent years, seen big advances because of the spread of internet, which generates everyday a tremendous volume of data, and also the immense advances in technologies which facilitate the analysis of these data. In particular, classification techniques are a subdomain of Data Mining which determines in which group each data instance is related within a given dataset. It is used to classify data into different classes according to desired criteria. Generally, a classification technique is either statistical or machine learning. Each type of these techniques has its own limits. Nowadays, current data are becoming increasingly heterogeneous; consequently, current classification techniques are encountering many difficulties. This paper defines new measure functions to quantify the resemblance between instances and then combines them in a new approach which is different from actual algorithms by its reliability computations. Results of the proposed approach exceeded most common classification techniques with an f-measure exceeding 97% on the IRIS Dataset.

Keywords—Data mining, knowledge discovery, machine learning, similarity measurement, supervised classification.

I. INTRODUCTION

Data Mining is a recent field at the cross junction of statistics and artificial intelligence, its purpose is the discovery of structures and information in large datasets. It consists of analyzing this profusion by using different algorithms which will explore the most infinite relation between these data and then deliver nontrivial, previously unknown and potentially useful information. Data mining was presented in 2001 by the Massachusetts Institute of Technology as one of the 10 emerging technologies that are changing the world in the 21st century [1].

In this dynamic environment, characterized by both technological advances on the one hand and the huge quantity of generated data on the second, analytical needs have also evolved, requiring the development of advanced algorithms and tools increasingly powerful which should be capable of adapting the contextual information to the types of data processed [2].

In general, Data Mining consists of solving a problem related to the considered data. There are four main issues addressed by this step:
- Association: Formally, it consists in discovering the most frequent relationships between attributes [3].
- Prediction: it seeks to estimate the value of a target attribute based on the values of the available attributes.

Depending on the type of the target variable, we distinguish: classification when predicting discrete values and regression when predicting numerical values [4].
- Grouping: it aims to generating a partition, consisting of a set of clusters where the data of the same cluster are very similar [5].
- Detection of anomalies: it detects aberrant behaviors in supposedly homogeneous data [6].

These problems generate two types of Data Mining model: predictive (supervised) and exploratory (unsupervised). Supervised learning consists of two major steps: training and testing [7].

The problem of classification is central in the majority of these applications. By its simplicity, it represents, in a way, the cornerstone of many treatments and gives rise to a variety of applications. It concerns business [8], medicine [9], finance [10], text analysis [11], and image analysis [12], etc. As a result, it has generated the most data mining work, so a large number of algorithms in the literature are devoted to the classification of textual data, which makes their complete presentation impossible. The most widely used supervised algorithms in the literature are Naive Bayes (NB) [13], Support Vector Machines (SVM) [14], k-Nearest Neighbors (KNN) [15] and Artificial Neural Networks (ANN) [16].

The goal of this study is to build up a new accurate statistical approach (CSBS) that supports different types of data, and that learns to classify from most reliable attributes more accurately.

In general, Data Mining consists of solving a problem related to the considered data. There are four main issues addressed by this step:
- Association: Formally, it consists in discovering the most frequent relationships between attributes [3].
- Prediction: it seeks to estimate the value of a target attribute based on the values of the available attributes.

These problems generate two types of Data Mining model: predictive (supervised) and exploratory (unsupervised). Supervised learning consists of two major steps: training and testing [7].

The problem of classification is central in the majority of these applications. By its simplicity, it represents, in a way, the cornerstone of many treatments and gives rise to a variety of applications. It concerns business [8], medicine [9], finance [10], text analysis [11], and image analysis [12], etc. As a result, it has generated the most data mining work, so a large number of algorithms in the literature are devoted to the classification of textual data, which makes their complete presentation impossible. The most widely used supervised algorithms in the literature are Naive Bayes (NB) [13], Support Vector Machines (SVM) [14], k-Nearest Neighbors (KNN) [15] and Artificial Neural Networks (ANN) [16].

The goal of this study is to build up a new accurate statistical approach (CSBS) that supports different types of data, and that learns to classify from most reliable attributes more accurately.

The rest of this paper is structured as follows: Section II summarizes main works on supervised machine learning. Section III meticulously details the modeling process, as it is crucial in the build of the CSBS decision function; and in Section IV, the results of four most known classifiers are compared to the CSBS. Finally, the last section concludes this work.

II. BACKGROUND

Until the end of the 1990s, the dominant approach to construct prediction models in general, and classification in particular, was based on a knowledge engineering approach. An expert system was constructed with a set of rules defined manually in order to emulate the decision-making ability of a human expert [17]. Expert systems were among the first truly successful forms of artificial intelligence [18]. However, the development of artificial intelligence techniques and computer performances has allowed the process of classification from another perspective. Rather than resorting to experts to establish the rules, the alternative came from a subdomain of...
artificial intelligence that developed strongly during the 1990s: machine learning that allows the system to learn by itself to accomplish this task [19]. Since then, several algorithms have succeeded:

A. K-Nearest Neighbors

The algorithm KNN is among the simplest artificial learning algorithms based on similarities. The basic idea when classifying a given instance is to vote its nearest neighbors in the sense of a predefined distance. The class of the new instance is then determined by the majority among the KNN.

The performance of KNN depends largely on two factors: the value of k and the measure used [20]. Generally, it becomes poor when the features are too many, or when data are very heterogeneous.

B. Naïve Bayes

NB classifier is based on Bayes’ theorem [21]. The basic idea is to estimate the probabilities of membership of the samples to their classes. The naive part of this model is the assumption that all variables are independent. The simplicity of this assumption makes the computation of NB classifier far more efficient [22].

C. Artificial Neural Network

This method is very schematically inspired from the function of biological neurons; hence its name. The neurons receive the signals (electrical impulses) through highly branched extensions of their cellular body (dendrites) and they send the information through long extensions (axons) [23].

The algorithm ANN learns a model by means of a feed-forward neural network trained by a back propagation algorithm. A neuron is primarily a mathematical operator. It performs a weighted sum, followed by a nonlinear function. This function must be bounded, continuous and differentiable; the most frequently used ones are sigmoid functions [24].

D. Support Vector Machines

The foundations of SVM originated from early concepts developed by Cortes and Vapnik [14], this method has proven to be very robust for general classification and regression [25].

SVM are based on two key ideas: the notion of maximum margin and the concept of kernel function. In linear classification, they create a hyper plan that separates the data into two sets with a maximum margin [26]. For the cases where the data are not linearly separable, they map the data representation space into an area of larger dimension in which it is probable that there is a linear separator.

Further works combined two or more of the previous algorithms. Machhale et al. proposed a classification system to recognize normal and abnormal MRI brain images by combining SVM and KNN [27]. Another work proposed a hybrid method based on ANN and KNN in a scheme intended for cloud classification [28]. In a work on opinion mining, noting that some unclassified instances by SVM could be done by KNN, authors have hybridized the SVM by KNN, thus raising the overall performance [29].

The optimization of previous algorithms also aroused the interest of researchers, Zhang et al. proposed a hybrid method applying SVM with an optimization of its parameters in an application on fault error diagnosis [30]. ANN and KNN algorithms have also been optimized to meet the need for an accurate classification [31], [32]. This diversity of algorithms is now applied in different domains:

The most remarkable interest was devoted to Customer relationship management (CRM) [33]. It comprised a set of processes and enabled systems supporting a business strategy to build long term, profitable relationships with specific customers [34]. The choices of data mining techniques should be based on the data characteristics and business requirements [35].

The field of healthcare also benefited from data mining algorithms; several works in the literature have applied supervised and unsupervised classification techniques whether to diagnose tumors [36], or predict the causes of a given disease [37], and many other clinical applications [38].

Classification techniques have also been applied to textual data, whether for automatic categorization of texts [39], or sentiment analysis [40], or simply for information retrieval [41].

Nowadays, these algorithms are applied in many other fields, and the need to construct more reliable approaches is becoming more and more indispensable [42].

III. THE PROPOSED APPROACH

In this section, the main problems faced by previous machine learning algorithms are highlighted, and new measures are defined in order to build up the new accurate approach we called CSBS.

The first faced problem while computing similarities for many attributes is the normalization:

Let’s consider the two examples: E₁ which belongs to the first class a, and E₂ which belongs to the second class b:

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>EXAMPLE OF NON-STANDARDIZED DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>E₁</td>
<td>3000 30 36 12 a</td>
</tr>
<tr>
<td>E₂</td>
<td>1200 55 64 21 b</td>
</tr>
<tr>
<td>E₃</td>
<td>2800 55 62 20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>DISTANCES TO CLASSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>a</td>
</tr>
<tr>
<td>Manhattan</td>
<td>64.75</td>
</tr>
<tr>
<td>Euclidean</td>
<td>101.69</td>
</tr>
</tbody>
</table>

When the distance is calculated directly, the impact of the order of magnitude is so important. For example, the new example E₃ has very close values to the example E₂ on the three attributes: x₁, x₃ and x₄; but the order of magnitude of x₁ (10^3) makes the entire distance very far, favoring thus the example of the class a which has relatively close value only for x₁; In order to overcome this problem, a normalization is adopted for each value eᵢ of the attribute xᵢ for the example Eᵢ.
The own amplitude: $A^{*}_{jk}$

The computation of the amplitude is used to express the reliability of each attribute:

Let’s reconsider the example of Table III, for a given example $E_{11}$, the attribute $x_1$ defines three possible cases according to its value:

- If $e_{11,1} \in [0.8 ; 1]$: $E_{11}$ has a big chance to belong to the class $b$ as it is the only class which takes values between 0.8 and 1.
- If $e_{11,1} \in [0.2 ; 0.8]$: It is impossible to know exactly if $E_{11}$ belongs to the class $b$ or $c$, as both of them include examples with values in this interval.
- If $e_{11,1} \in [0.1 ; 0.2]$: All classes have values in this interval; hence, the need to compute a new measure of similarity.
- If $e_{11,1} \in [0 ; 0.1]$: $E_{11}$ has a big chance to belong to the class $c$.

D. Own Amplitude $A^{*}_{jk}$:

The own amplitude of a given attribute is introduced to predict if this latter is reliable relatively to other attributes (see Fig. 1). It eliminates the intervals with uncertain values.

\[
A^{*}_{jk} = 0.2 - 0.1 = 0.1
\]

For example, for the class $a$, the attribute $x_1$ has the following characteristics:

- The number of examples: $n_k = 3$
- The minimal value: 0.1
- The maximal value: 0.2
- The center of the class: $c_{1a} = \frac{0.1+0.2+0.2}{3} = 0.17$
- The amplitude: $A_{1a} = 0.2 - 0.1 = 0.1$
- The own amplitude: $A^{*}_{1a} = 0$ because the class $c$ also has values in this interval (Fig. 1).

In what follows, different sub measures of similarities are discussed:

A. Equal Values $n_k$

While classifying $E_{10}$, let us begin with the attribute $x_4$: it has the same value as two examples from the class $a$ (66%). Similarly, $x_2$ has one equal value belonging to the same class $a$.

By adding the rest of attributes, the class of $E_{10}$ is the one having most equal values on all attributes. This measure is very valuable especially for discrete attributes, and it is used by Decision Tree algorithms to compute the information gain. However, being limited to this measure alone is very insufficient as it can be clearly seen for the attribute $x_3$ for instance, which has one equal value from the class $c$; but the two other examples are so far, and contrariwise, the values of the class $a$ are very close to it; hence, the need to consider the concentration of values around the center of the class (especially for continuous attributes).

B. Distance to Center $d_c$

To overcome the aforementioned problem of concentration of values, distances to centers can be envisaged.

The attribute $x_3$ which was close to the values of the class $a$ without being exactly equal to, has now the nearest value to its center:

\[
\text{Distance to } x_3 = 0.01
\]

\[
\text{Distance to center } d_c = 0.35
\]

\[
\text{Distance to center of } c = 0.52
\]

C. Amplitude $a_{jk}$

TABLE V

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{10}$</td>
<td>0.3</td>
<td>0</td>
<td>0.83</td>
</tr>
</tbody>
</table>

The goal is to build up a model that would be able to classify correctly any given new observation:

\[
e_{ij} \leftarrow \frac{e_{ij} - \min(e_{ij})}{\max(e_{ij}) - \min(e_{ij})}
\]
E. Decision Function

To classify a given instance \( E_t \), for each class \( k \), the proposed approach computes the likelihood, which is defined as:

\[
\xi_{lk} = \frac{1}{c_k} \times \sum_{j=1}^{N_k} \frac{N_j^{k} + a_k^{k} + a_k^{k} - d(x_{ij}, x_j^k)}{N_j + 2a_k^{k} + \epsilon}
\]

where: \( a_k^{k} \) denotes the amplitude of the interval of values of \( k \), \( a_k^{k} \) denotes the own amplitude of the interval of values of \( k \), \( N_j^{k} \) is the number of instances of the class \( k \) having a value equal to \( E_t \) on the attribute \( x_j \), \( N_j \) is the total number of instances having a value equal to \( E_t \) on \( x_j \), \( d(x_{ij}, x_j^k) \) denotes the considered distance between \( E_t \) and the center \( x_j^k \), \( p_{jk} \) is the coefficient of reliability on \( x_j \) to predict the class \( k \), \( c_k \) is the number of instances of the class \( k \), and, \( \epsilon \) is a very small positive value added to avoid a zero denominator (where all instances take the same value for \( x_j \), and this latter is different from that of \( E_t \)). The numerator remains zero for this case. Finally, the class of \( E_t \) is:

\[
Y_i = \text{argmax}_k \xi_{lk}
\]

A threshold \( \delta \) is chosen to eliminate non-significant attributes (attributes with smallest reliability coefficients). Such a threshold greatly increases the precision of the classification since it relies solely on the attributes which separates the classes, but increasing it leads to over-classification.

IV. RESULTS AND DISCUSSION

To evaluate the performance of the proposed approach, the effectiveness is compared to most common approaches of classification: k-NN, ANN, SVM, NB, on IRIS dataset which is a classic example of datasets for classification [42].

A. Computation

The dataset was divided into five subsets of 30 instances each (10 instances per class: I. setosa; I. versicolor; I. virginica).

At each iteration, four subsets are considered for training and the fifth for the test, the process is thus repeated for the five subsets, and the average f-measure is retained.

The F-Measure used for evaluation is written as:

\[
\text{Recall} = \frac{a}{a+c} ; \text{Precision} = \frac{a}{a+b} ; \text{FM} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}
\]

where \( a \) is the number of instances correctly classified, \( b \) : the number of false positives, and \( c \) : is the number of false negatives.

In what follows, the results of classification of the proposed algorithm CSBS and four other known algorithms:

B. K-Nearest Neighbors

Fig. 2 shows the f-measure of KNN algorithm for different values of \( k \). The optimal performance concerns \( k = 3 \), and the f-measures of the five subsets are: 100%, 96.82%, 93.89%, 96.82% and 96.82%, with a total of five instances erroneously classified.

C. Naïve Bayes

In Fig. 3, the NB algorithm generates the following results on the five subsets: 100%, 93.89%, 93.89%, 96.82% and 93.33%. The average f-measure is: 95.58%.

D. Artificial Neural Network

Fig. 4 summarizes the f-measure of ANN on the five subsets: 100%, 93.33%, 91.14%, 100% and 100%.

E. Support Vector Machines

In Fig. 5, the results concern SVM with RBF kernel [43] \( y = 0, C = 0 \) and \( \epsilon = 10^{-3} \). On the five subsets, the reported f-measures are: 100%, 93.33%, 90.12%, 96.28% and 90.12%.

F. CSBS

The f-measure of CSBS algorithm showed a stability on the five subsets: 100%, 96.82%, 93.89%, 100%, 96.82%. None of the subsets generated more than three erroneous instances. The average f-measure is the highest among previous
algorithms: 97.50%.

The overall results are shown in Fig. 7 in terms of f-measure: the algorithm CSBS slightly exceeds KNN (k = 3) with the highest average f-measure: 97.5%. ANN returned the third best f-measure: 96.89%, and finally, NB and SVM (with a RBF kernel function) returned the weakest f-measures: respectively 95.58% and 94.07%. Moreover, the f-measure of CSBS algorithm is relatively stable on the five subsets. The weakest reported f-measure was 93.89%, and it concerns the training set with the lowest reliability coefficients, which means that when the attributes are sufficiently reliable, the CSBS algorithm outperforms other algorithms; But when this is not the case, some algorithms may offer better results, such as ANN on the subset s5. Future improvements can then be considered to hybridize CSBS with other supervised classification algorithms, for example with ANN, in order to make it even more efficient.

The outperforming of the CSBS algorithm is even more remarkable when the data type is more heterogeneous (for example on a sentiment analysis dataset).

V. CONCLUSION AND PERSPECTIVES

Data mining is a fast expanding field with many new research results and new approaches developed. Applications of data mining techniques affect almost every aspect of our daily lives: industry, medicine, and finance, etc., and it has attracted the attention of practitioners and academics.

Nowadays, as the world is witnessing the increasingly complex needs of big data, classic algorithms are pushed to their limits, and new accurate machine learning approaches are necessary to deal with the variety of available data. In this sense, this paper introduces a new supervised machine learning approach that aims to overcome main drawbacks of classic classification models. This algorithm, called CSBS, selects most reliable attributes for each class, and then combines its reliability coefficients and special similarity measures to classify any given instance. A standard dataset is used to evaluate the proposed algorithm, and the results indicates that it outperforms most known classification algorithms, namely KNN, NB, ANN and SVM with a f-measure exceeding 97%. Afterwards, an analysis of the final results highlighted that CSBS can be further optimized by hybridization with ANN classifier. This optimization is the subject of our future researches.

REFERENCES


