A NOVEL METHODOLOGY FOR CONSTRUCTING RULE-BASED NAÏVE BAYESIAN CLASSIFIERS

Abdallah Alashqur

Faculty of Information Technology, Applied Science University, Amman, Jordan

ABSTRACT

Classification is an important data mining technique that is used by many applications. Several types of classifiers have been described in the research literature. Example classifiers are decision tree classifiers, rule-based classifiers, and neural networks classifiers. Another popular classification technique is naïve Bayesian classification. Naïve Bayesian classification is a probabilistic classification approach that uses Bayesian Theorem to predict the classes of unclassified records. A drawback of Naïve Bayesian Classification is that every time a new data record is to be classified, the entire dataset needs to be scanned in order to apply a set of equations that perform the classification. Scanning the dataset is normally a very costly step especially if the dataset is very large. To alleviate this problem, a new approach for using naïve Bayesian classification is introduced in this study. In this approach, a set of classification rules is constructed on top of naïve Bayesian classifier. Hence we call this approach Rule-based Naïve Bayesian Classifier (RNBC). In RNBC, the dataset is canned only once, off-line, at the time of building the classification rule set. Subsequent scanning of the dataset, is avoided. Furthermore, this study introduces a simple three-step methodology for constructing the classification rule set.

Keywords

Data Mining, Classification, Bayes Theorem, Rule-Based Systems, Machine Learning.

1. INTRODUCTION

In many applications, it is important to classify data stored in a dataset. Each record in the dataset needs to be associated with a certain class. One column in the data set (normally identified as the class label) stores the class name for each record. Typically there are very few classes, where each class is shared by many records in the dataset. The aim of classification asa data mining technique is to be able to *automatically* classify new records whose classes have not yet been determined [1,2]. In order for this to be achievable, the data mining system has to have the knowledge based on which it can classify new records. This is normally done by supplying the data mining system with apre-classified dataset from which it can derive (i.e., learn or infer) the criteria used to determine the classes of data records in this dataset. After the system learns the classification criteria, it can use them to predict the classification of new unclassified records [1,3,4].

In data mining terminology, the classification criteria are sometimes referred to as classification model. The classification model is a representation scheme used to capture and represent the classification criteria. In addition, the pre-classified dataset, from which the system learns, is referred to as *training* set since its sole purpose is to train the system. Therefore, the classification process goes through two phases. The first phase is the learning phase-in which the classification model is derived from the training set.Once the system learns the classification model, and after going through some testing, the system is ready for the second phase. In the second phase, which

DOI:10.5121/ijcsit.2015.7114

we call the application phase, the system applies the classification model to new records in order to infer or predict their classes.

Typically the training set consists of historical data whose classes have become known. A training set can also be constructed by giving unclassified data to experts in the application domain who can classify the records manually based on their expert knowledge. In any way, the level of trust in the correctness and integrity of classificationsin the training set should be very high, since it is used to train the system during the learning phase.

Because we provide the classification system with a training set whose records have been classified, classification is considered a type of *supervised* learning. This distinguishes it from other techniques such as clustering, in which no training set is provided to the system. In Clustering, the data mining system is supposed to logically partition the dataset into clusters (similar to classes) on its own without learning from a pre-clustered dataset. Therefore, Clustering is referred to as *unsupervised* learning.

Normally a classification technique is used to derive and capture the classification model. Several classification techniques have been presented in the literature. Some examples of these techniques include decision tree classifiers [5,6,7,8], rule-based classifiers [9,10], artificial neural networks [11], Bayesian classifiers [11,12], support vector machines (SVM) [13,14]and ensemble methods [15]. One of the main differences between these algorithms is the way the learned classification model is represented. For instance, in decision tree induction, the derived model is represented in the form of a decision tree. A rule-based classifier, on the other hand, represents the learned model in the form of a collection of If-Then rules. Bayesian classifiers are statistical classifiers in which the learned model is embodied in a set of equations based on Bayes' theorem.

The focus of this study, which is part of an on-going research project called Probabilistic Data Management and Mining (PDMM), is on the learning phase of the classification process in which Naïve Bayesian classification is used. In naïve Bayesian classification, whenever a new record is to be classified, the entire dataset needs to be scanned to gather statistics and apply a set of statistical equations. The outcome of these equations is a probabilistic prediction of the class of the newly inserted record. The need to scan the entire dataset every time a new record is inserted is considered a problem especially if the dataset is very large. This is due to the high cost of the scanning step. To alleviate this problem, a new approach for using naïve Bayesian classification is introduced in this study. This approach uses a probabilistic model that is based on Naïve Bayesian classification for building a set of classification rules, hence it is called Rule-based Naïve Bayesian Classifier(RNBC). In RNBC, the Bayesian statistical equations are applied to the dataset only at the beginning in order to derive (or compile) a set of classification rules that cover all possible cases. From that point on, whenever a new record is to be classified, the set of classification rules is searched to find the rule that is *satisfied* by the record. That rule is then fired (i.e., applied to the record) to derive the record's classification. This way the Bayesian equations don't have to be evaluated against a large dataset every time a newly inserted record is to be classified. Furthermore, in this study we introduce a three-step methodology for building such a rule-based classifier.

2. RELATED BACKGROUND

RNBC can be considered as a combination of two distinct classification approaches, namely, naïve Bayesian classification and rule-based classification. In this section we provide a brief

background on both of these approaches. In subsequent sections, we describe RNBC and the three-step methodology used to build it.

Broadly speaking, there are two generic approaches for building a rule-based classifier. They are summarized as follows.

1. Direct Approach. In this approach, the classification rules are learned *directly* from the training dataset. Sequential covering algorithms are of this type. In Sequential covering algorithms, rules are extracted sequentially, i.e., one at a time. Each time a rule is extracted, the records covered by the rule are removed from the dataset, and the process is repeated on the remaining records. An example algorithm that follows this approach is RIPPER [16].
2. Indirect Approach.In this approach, the rules are not learned directly from the data set, but they are derived from another classification technique such as decision trees. The popular C4.5 classification algorithm falls in this category [1].In C4.5 algorithm, a decision tree is learned first, then the set of classification rules are derived from the decision tree.

Research presented in this study falls in the second category above. In RNBC, Naïve Bayesian classification is used as a step towards building a rule-based classifier. Therefore, RNBC inherits the benefits of both naïve Bayesian classification as well as rule-based classification. From Naïve Bayesian classification,RNBC inherits the high degree of classification accuracy. Whereas the rule-based approach gives RNBC the ability to classify new records without having to frequently scan the dataset.

2.1 Rule-Based Classification

We demonstrate how a rule-based classifier works by using the training set shown in Table 1 for a financial institution. This training set is pre-classified and the class label is *May Default*, which identifies who may default on their loan if they borrow money from the financial institution. There are two classes in this dataset: "YES" and "NO".From this data set, the system learns the set of rules that can be used to classify newly-inserted records. The rules are of the form:

R_x: $(Condition_1 \wedge Condition_2 \wedge \ldots \wedge Condition_n) \rightarrow Class$

Where R_x is the rule-id andthe left hand side (LHS) of the rule represents *conditions* on some or all of the attributes in the dataset except the class label. The right hand side (RHS) is the*class name*, which in our example is either "YES" or "NO".An example classification rule based on the dataset of Table 1 is:

R_A:$(Annual\ Income = high) \wedge (Home\ Owner = NO) \wedge (Age \geq 40) \rightarrow (May\ Default = NO)$

A rule R is said to *cover* a record if the attributes of the record satisfy the conditions of R. Hence rule R_A covers record 2 and record 9 in the dataset of Table 1. Also a record is said to *satisfy* a rule if all the conditions in the LHS of the rule are true for that record.The coverage of a rule is defined as the percentage of records in the dataset that satisfy the rule [11], which can be formulated as follows.

$$Coverage(R) = \frac{Numberofrecordsthatsatisfy R}{Totalnumberofrecords}$$

Based on this equation, the coverage of rule R_A is 2/10 = 0.2.

Table 1: A classified Dataset

ID	Annual Income	Home Owner	Age	May Default
1	High	Yes	< 40	No
2	High	No	≥ 40	No
3	High	No	< 40	YES
4	Medium	Yes	≥ 40	No
5	low	No	≥ 40	No
6	Medium	No	≥ 40	No
7	Medium	Yes	≥ 40	Yes
8	Medium	No	< 40	Yes
9	High	No	≥ 40	No
10	Low	No	< 40	Yes

Rules can be *mutually exclusive* if no record satisfies more than one rule. The set of rules can be*exhaustive* if the rule set covers the entire dataset, i.e., every record is covered by at least one rule. When a record is satisfied by more than one rule, normally some conflict resolution techniques are applied to determine which rule to apply (Tan et al., 2014). One of these techniques is to assign priorities to the rules, and the rule with the highest priority is applied.

2.2 Naïve Bayesian Classification

Naïve Bayesian classification [11,12)is based on Bayesian Theorem. If a new record R is to be classified, Bayesian Theorem can be used to find the probability that it belongs to class C_i by using Eq. (1) shown below.

$$P(C_i|R) = \frac{P(R|C_i)P(C_i)}{P(R)} \quad (1)$$

Where P denotes *probability* and the notation*P(X|Y)* represents the conditional probability of X given that Y has occurred.$C_i$is one of a set of classes *{ C1, C2, C3, …}* that are used to classify the data. For example, in Table 1 there are two classes as determined by the attribute *May Default*, namely *{YES, NO}*. Equation (1) is computed for every class C_i. The class whose *P(Ci |R)* is highest is selected as the record's class.

When computing$P(C_i|R)$ for every $C_i$to determine the class with the highest probability, the denominator *P(R)* is constant across all classes. Therefor it can be removed fromthe computations. Therefore Eq. (2) below can be used to find the class with the highest probability.

$$P(C_i|R) \sim P(R|C_i)P(C_i) \quad (2)$$

Where the symbol "~" indicates that the LHS is *proportional* to the RHS.Further, Naïve Bayesian classification assumes *class-conditional independence* (that is why it is called "naïve"). This assumption basically states that attribute values of the record *R*are independent of each other. In other words, if *R* is the n-record$<r_1, r_2, \ldots r_n>$, then *P(R|Ci)* in Equation (2) can be computed as shown in Equation (3) below.

$$P(R|C_i) = \prod_{k=1}^{n} P(r_k|C_i) = P(r_1|C_i) \times P(r_2|C_i) \times \ldots \times P(r_n|C_i) \quad (3)$$

This is because, from Probability Theory, the probability of the conjunction of *independent* events can be obtained by multiplying the probabilities of the individual events. In summary, to compute $P(C_i|R)$ based on Eq (2) we need to compute $P(R|C_i)$ based on Eq. (3)and compute $P(C_i)$ then multiply the two results. This needs to be repeated for each class C_i.

2.3 Bayesian Classification Example

Assume we have a new record *R = < LOW, YES, "<40">*for Table 1where the record's values are for the attributes: *Annual Income, Home Owner*, and *Age*, in that order. We ignore the *ID* attribute since it is just used as a primary key for identification purposes. We want to use Naïve Bayesian Classification to predict the class *May Default*(i.e. whether it is YES or NO) for this record. Based on Eq. (2), we need to find the values of the terms on the right hand side of the equation, namely $P(C_i)$ and $P(R|C_i)$. First we compute $P(C_i)$ for all classes. In other words, we need to find *P(May Default = YES)* and *P(May Default = NO)*. In Table 1, out of ten records, there are four records whose class is YES and there are six records with class NO. Therefore the probabilities of these two classes are as shown below.

P(May Default = YES) = 4/10= 0.4(4)
P(May Default = NO) = 6/10 = 0.6(5)

To compute the conditional probability $P(R|C_i)$ we use Eq. (3). We need to find $P(r_k|C_i)$ for each r_k and for each C_i where r_k represents attribute values for the record *R = <"LOW", "YES", "<40">*. This is performed below.

Computing $P(R|C_i)$ forClass (*May Default = YES*). Since there are four records with class May default = YES and only one of them has*Income = Low*, then

P(Income = LOW| May Default = YES) = 1/4= 0.25

Similarly the computations of the rest of the attribute values are as follows.

P(Home Owner = Yes| May Default = YES) = 1/4= 0.25
P(Age = "<40"| May Default = YES) = 2/4= 0.5
Substituting in Eq.3, we obtain
P(R | May Default = YES) = 0.25 x 0.25 x 0.5 = 0.03(6)

Computations$P(R|C_i)$ for class (*May Default = NO*). These probabilities are computed as follows.

P(Income = LOW| May Default = NO) = 1/6= 0.17
P(Home Owner = Yes| May Default = NO) = 2/6= 0.34
P(Age = "<40"| May Default = NO) = 1/6= 0.17

Substituting in Eq.3, we obtain:

P(R | May Default = NO) = 0.17 x 0.34 x 0.17 = 0.01(7)

To apply Eq.2 for class *(May Default = YES)* we need to multiply the results of equations (4) and (6) to obtain:

P(May Default = YES| R)~0.4 X 0.03 = 0.012(8)

Now we can apply Eq. (2) for class *(May Default = NO)*bymultiplying the results of equations (5) and (7) to obtain:

P(May Default = NO| R)~ 0.6 X 0.01 = 0.06(9)

To compute the exact probabilities of*P(May Default = YES | R)* and*P(May Default = NO | R)* instead of the proportionality ("~") shown in Eq. (8) and (9), we can do so be observing, from Probability Theory, that the sum of the two probabilities in these equations should add up to one.

Therefore,

P(May Default = YES | R) + P(May Default = NO | R) =1

By substituting the two results of Eq. (8) and (9) and multiplying by a normalizing factor N, one can obtain:

*N × (0.012 + 0.06) = 1.*Solving for N we get:
N = 13.889

Therefore,

P(May Default = YES | R) = 0.012 x N = 0.012 x 13.889 = 0.1667

And similarly for P(May Default = NO | R)as shown below.

P(May Default = NO | R)= 0.06 x 13.889= 0.8333

By comparing the values of *P(May Default = YES | R)* and P(May Default = YES | R) it is clear that*P(May Default = NO | R)* is larger than *P(May Default = YES | R).* Thereforeone concludes that the new record *R = <LOW,YES, "<40">* should be classified as *May Default = NO.*

The above computation process is repeated for every new record in order to predict its classification.

3. THREE-STEP METHODOLOGY FOR BUILDING RNBC

In this section, a three-step methodology for building a rule-based classification system that is based on Bayesian classification is introduced. In in this approach, there is a learning phase in which the system follows the three steps to extract classification rules. Note that in this approach and in classification in general, the attributes are assumed to be discretized (or categorized). If the values in these attributes are continuous, a pre-processing phase is performed to discretize them.

3.1 Description of the Methodology

The steps of the methodology used in RNBC are outlined below.

Step 1.Generate all possible combinations of attribute values that exist in the dataset.
Step 2.For each combination of attribute values found in step 1, compute the probability of each class.
Step 3.Generate the classification rules, one rule for each combination of attribute values found in Step 1. The class designated by each rule is the class with the highest probability as found in Step 2.

An example against the dataset shown in Table 2 is used to demonstrate how these steps are performed and the outcome of each step. This dataset has been extracted from our earlier research

on association rule mining[17,18].The dataset of Table 2 contains data pertaining to ex-members of a gym club. In other words, this is historical data that is stored in the database for members who terminated their membership. This data includes AGE (A), GENDER (G), MEMBERSHIP_DURATION (MD) to represent how long a member maintained a valid membership in the club, HOME_DISTANCE (HD) to represent how far a member's residence is from the club location, and HOW_INTRODUCED (HI) to represent how a member was originally introduced to the club such as by referral or by seeing an advertisement in a newspaper. Table 2 shows this dataset as populated with sample data. In real life situations, a large club, with many branches, may have millions of ex-members, thus millions of records may exist in such a dataset.

Table 2.Sample Dataset

ID	AGE(A)	GENDER(G)	HOME DISTANCE (HD)	HOW INTRODUCED (HI)	MEMBERSHIP DURATION (MD)
1	young	F	close	newspaper	short
2	middle	M	far	newspaper	short
3	senior	F	close	referral	long
4	senior	F	close	referral	long
5	young	F	far	newspaper	long
6	middle	M	close	newspaper	short
7	senior	M	far	newspaper	short
8	senior	F	close	referral	long
9	young	F	close	referral	long
10	middle	F	far	newspaper	long
11	middle	M	far	newspaper	short
12	senior	F	close	referral	long
13	senior	M	far	referral	short
14	middle	M	far	referral	long

The two classes that exist in the MD column are "short" and "long". A classification system learns the classification model from this dataset in order to be able to predict whether a new member is expected to stay as member for a long period or for a short period. This knowledge is useful from a business perspective since the club's management may offer incentives directed to those who are expected to stay for a short period to encourage them to renew their memberships.

3.2 Applying the methodology to an Example

In what follows we demonstrate how each step is performed and its outcome.

Step 1.In this step all possible combinations of attribute values are generated. We ignore the *ID* attribute since it is just used as a primary key. Also MEMBERSHIP DURATION (MD) is ignored in this step since it represents the class that is to be predicted in Step 3. To find all possible combinations of attribute values, we fist identify the set of distinct allowable values for each attributes (i.e., the domain of the attribute) as follows.

AGE = {young, middle, senior}

GENDER = {f, m}
HOME DISTANCE = {far, close}
HOW INTRODUCED = {referral, newspaper}

We assume that each attribute is limited to these values by a constraint on the database. The alternative is that the transactional dataset contains other more continuous values, but that dataset is preprocessed and converted to the dataset show in Table 2 by using *discretization* (or *categorization*), which is a popular preprocessing technique in data mining.

The records that represent all possible combinations of values can be found by taking the *cross product* of the above sets. The number of these records can be found by multiplying the number of values in each set as follows.

Number of records with unique possible combinations =
$|AGE| \times |GENDER| \times |HOME\ DISTANCE| \times |HOW\ INTRODUCED| =$
$3 \times 2 \times 2 \times 2 = 24$

We use |AGE| to denote the number of values in the set for Age. The same goes for other attributes. The output of Step 1 is shown in Table 3.

Table 3. All possible unique combinations of attribute values

ID	A	G	HD	HI
1	young	f	close	newspaper
2	young	f	close	Refer
3	young	f	far	newspaper
4	young	f	far	referral
5	young	m	close	newspaper
6	young	m	close	referral
7	young	m	far	newspaper
8	young	m	far	referral
9	middle	f	close	newspaper
10	middle	f	close	referral
11	middle	f	far	newspaper
12	middle	f	far	referral
13	middle	m	close	newspaper
14	middle	m	close	referral
15	middle	m	far	newspaper
16	middle	m	far	referral
17	senior	f	close	newspaper
18	senior	f	close	referral
19	senior	f	far	newspaper
20	senior	f	far	referral
21	senior	m	close	newspaper
22	senior	m	close	referral
23	senior	m	far	newspaper
24	senior	m	far	referral

Step 2. In this step, one computes the probability of whether the record is classified as (*MEMBERSHIP_DURATION = long)* or as *(MEMBERSHIP_DURATION = short)* for each of the twenty four records shown in Table 3. Below we compute these probabilities for the first record of Table 3.

As explained in Section 2, Equations (2) and (3) need to be applied to compute the class probability. First we compute the probabilities of the two classes *long* and *short*. In other words, what is needed is to find the two probabilities *P(MEMBERSHIP DURATION = long)*and *P(MEMBERSHIP DURATION = short)*. As a short notation we use *P(long)* and *P(short)* respectively to denote *P(MEMBERSHIP DURATION = long)*and *P(MEMBERSHIP DURATION = short).*

The dataset of Table 2 is used as a basis to compute *P(long)* and *P(short)* since it is the training set. Out of 14 records, there are 8 records whose *MD = long* and 6 records whose *MD = short*. Hence the probabilities are:

P(long)= 8/14 = 0.571(9)
P(short)= 6/14= 0.429(10)

Next Equation (3) is used to compute *P(R|long)* and *P(R|short)*, where *R* is the first record in Table 3. To compute *P(R|long)*we need to compute $P(r_i|long)$ for each r_i, where r_i is a component value of *R*.

Computing *P(R|Long).*Given that the components of R, the first record in Table 3, are represented by the list<young, f, close, news-paper>, one can compute *P(R|long)* as follows.

P(AGE = YOUNG | long)= 2/8=0.250
P(GENDER= F |long)= 7/8=0. 875
P(HOME_DESTANCE = CLOSE | long)=5/8=0.625
P(HOW_INTRODUCED = NEWS_PAPER| long)= 2/8=0.250

Now Equation (3) is applied in order to compute P(R|long) by multiplying the probabilities above to obtain:

P(R|long) = 0.250 x 0. 875x 0.625 x 0.250 =0.0342(11)

Computing P(R | short). Similarly P(R|short) can be computed as follows.

P(AGE = YOUNG| short)= 1/6=0.166
P(GENDER = F | short)= 1/6=1.66
P(HOME_DESTANCE = CLOSE| short)=2/6=0.333
P(HOW_INTRODUCED =NEWS_PAPER|short)= 5/6= 0.833

Multiplying to get *P(R|short):*

P(R|short) = 0.166 x 0.166 x 0.333 x 0.833= 0.00764(12)
Now the results of Equations (9) and (11)are substituted into Equation (2) to obtain P(long|R).
P(long|R)~0.0342×0.571 =0.0193
And substituting from Eq. (10), (12) into Eq. (2) to obtainP(short | R).:
P(short | R)~0.00764×0.429= 0.00328.
We can find the exact probabilities for *P(long | R)* and *P(short | R)* by noting that, from Probability Theory, the two values above should add up to one. Hence,

$P(long \mid R) + P(short \mid R) = 1$

Or

$N \times (0.0193 + 0.00328) = 1$

Where N is a normalizing factor. Solving for N, one obtains:

$N = 44.3$

Therefore,

$P(long \mid R) = N \times 0.0193 = 44.3 \times 0.0193 = 0.85$
$P(short \mid R) = N \times 0.00328 = 44.3 \times 0.00328 = 0.15$
Since $P(short \mid R) < P(long \mid R)$, the classification of this record can be assumed to be *Membership Duration = long*.

In a similar way, the probabilities of $P(long \mid R)$ and $P(short \mid R)$ are computed for the rest of the records in Table 3 as part of Step 2 of our methodology. The resulting values of $P(long \mid R)$ and $P(short \mid R)$ are shown in Table 4 for all records of Table 3.

Table 4. Computation Results for All instances

ID	A	G	HD	HI	P(long)	P(short)
1	young	f	close	news-paper	0.85	0.15
2	young	f	close	referral	0.99	0.01
3	young	f	far	news-paper	0.47	0.53
4	young	f	far	referral	0.9	0.1
5	young	m	close	news-paper	0.14	0.86
6	young	m	close	referral	0.72	0.28
7	young	m	far	news-paper	0.91	0.09
8	young	m	far	referral	0.6	0.4
9	middle	f	close	news-paper	0.66	0.34
10	middle	f	close	referral	0.03	0.97
11	middle	f	far	news-paper	0.63	0.37
12	middle	f	far	referral	0.1	0.9
13	middle	m	close	news-paper	0.95	0.05
14	middle	m	close	referral	0.54	0.46
15	middle	M	far	news-paper	0.98	0.02
16	middle	m	far	referral	0.8	0.2
17	senior	f	close	news-paper	0.14	0.86
18	senior	f	close	referral	0.01	0.99
19	senior	f	far	news-paper	0.36	0.64
20	senior	f	far	referral	0.04	0.96
21	senior	m	close	news-paper	0.86	0.14
22	senior	m	close	referral	0.28	0.72
23	senio	m	far	news-paper	0.95	0.05
24	senior	m	far	referral	0.57	0.43

Step 3. In this step all classification rules are created. Each record in Table 4 is examined to see which class has higher probability for that record. A classification rule is generated for each record. The THEN part of the rule designates the class with the highest probability and the IF part represents the conditions on all values. As an example, the class probabilities for the first record has the two probabilities: *P (MD = long) = 0.85* and *P(MD = short) = 0.15*. Since *P (MD = long)* is larger, it is used by the classification rule for the first record. Therefor the classification rule based on the first record of Table 4 can be formulated as follows.

(A =young)∧ (G = f)∧ (HD = close) ∧ (HI = news-paper) → (MD = long)

The rest of the classification rules are derived in a similar way. Figure 1 shows all 24 rules for all records in Table 4.

R01:(A =young) ∧ (G = f) ∧ (HD = close) ∧ (HI = news-paper) → (MD = long)

R02:(A =young) ∧ (G =f) ∧ (HD =close) ∧ (HI = referral) → (MD = long)
R03:(A =young) ∧ (G =f) ∧ (HD =far) ∧ (HI = news-paper) → (MD = short)

R04:(A =young) ∧ (G =f) ∧ (HD =far) ∧ (HI = referral) → (MD = long)

R05:(A =young) ∧ (G =m) ∧ (HD =close) ∧ (HI = news-paper) → (MD = short)

R06:(A =young) ∧ (G =m) ∧ (HD =close) ∧ (HI = referral) → (MD = long)
R07:(A =young) ∧ (G =m) ∧ (HD =far) ∧ (HI = news-paper) → (MD = long)

R08:(A =young) ∧ (G =m) ∧ (HD =far) ∧ (HI = referral) → (MD = long)

R09:(A =middle) ∧ (G =f) ∧ (HD =close) ∧ (HI = news-paper) → (MD = long)

R10:(A =middle) ∧ (G =f) ∧ (HD =close) ∧ (HI = referral) → (MD = short)
R11:(A =middle) ∧ (G =f) ∧ (HD =far) ∧ (HI = news-paper) → (MD = long)

R12:(A =middle) ∧ (G =f) ∧ (HD =far) ∧ (HI = referral) → (MD = short)

R13:(A =middle) ∧ (G =m) ∧ (HD =close) ∧ (HI = news-paper) → (MD = long)

R14:(A =middle) ∧ (G =m) ∧ (HD =close) ∧ (HI = referral) → (MD = long)

R15:(A =middle) ∧ (G =m) ∧ (HD =far) ∧ (HI = news-paper) → (MD = long)

R16:(A =middle) ∧ (G =m) ∧ (HD =far) ∧ (HI = referral) → (MD = long)

R17:(A =senior) ∧ (G =f) ∧ (HD =close) ∧ (HI = news-paper) → (MD = short)
R18:(A =senior) ∧ (G =f) ∧ (HD =close) ∧ (HI = referral) → (MD = short)
R19:(A =senior) ∧ (G =f) ∧ (HD =far) ∧ (HI = news-paper) → (MD = short)

R20:(A =senior) ∧ (G =f) ∧ (HD =far) ∧ (HI = referral) → (MD = short)

R21:(A =senior) ∧ (G =m) ∧ (HD =close) ∧ (HI = news-paper) → (MD = long)

R22:(A =senior) ∧ (G =m) ∧ (HD =close) ∧ (HI = referral) → (MD = short)
R23:(A =senior) ∧ (G =m) ∧ (HD =far) ∧ (HI = news-paper) → (MD = long)

R24:(A =senior) ∧ (G =m) ∧ (HD =far) ∧ (HI = referral) → (MD = long)

Figure 1: Learned Bayesian Classification Rules

The set of rules is exhaustive in the sense that any newly inserted record must be covered by one of the rules. After Step 3 is completed, we end up with a set of classification rules similar to those shown in Figure 1. From that point on, whenever a new record is to be added, the system does not have to scan the existing dataset in order to apply Bayesian equation and figure out the class. Instead, the data mining system finds the rule whose conditions are met by the data in the new record. That rule is *fired* to infer the class.

Refreshing the Rule Set. After adding several new records to the system, the probabilities of the values may differ from what they were before, and the existing classification rules may not reflect precisely what exists in the dataset. To solve this problem we propose to refresh the rules set periodically. The refreshing frequency may be set by the administrator of the system. For example, the administrator may instruct the system to refresh the rule set by re-computing the rulesevery time the dataset grows by 10%. This is done by applying the three-step methodology that we introduced in this paper (even though step 1 may not need to be repeated at refresh time). Periodically refreshing the rule set insures that the rules stay in-synch with the data. This will not degrade performance since the refreshing process can be performed off-line.

4. CONCLUSION

This study introduced a new approach for building a rule-based classifier called RNBC. In this approach, a rule-based classifier is built byderiving it froma Naïve Bayesian classifier. A three-step methodology for building such a rule-based classifier was also introduced. A detailed example is used to demonstrate how the steps of the methodology are applied to a dataset.

By using RNBC, whenever a new record is to be classified, the set of rules are searched to find the rule that applies. That rule is then fired to infer the record's class. The need to scan the dataset with each classification instance is avoided. In traditional naïve Bayesian classification, on the other hand, every time a new record is to be classified, the entire dataset needs to be scanned and a set of equations needs to be applied. Thus RNBC is considered an improvement over existing naïve Bayesian classifiers.

The set of classification rules in RNBC needs to be refreshed periodically as time progresses and the dataset grows, in order for the rules to stay up-to-date. But the refreshing process to build the updated set of rules can be performed off-line, therefore it won't impact the system's performance.

ACKNOWLEDGEMENT

The author is grateful to the Applied Science Private University, Amman, Jordan, for the full financial support granted to this research project (Grant No. DRGS-2014-2015-2).

REFERENCES

[1] Tan, P.N., 2014. Introduction to Data Mining. 2nd Edn., Pearson Education Ltd., UK., ISBN10: 0133128903.

[2] Fu, Z., B.L. Golden, S. Lele, S. Raghavan and E. Wasil, 2006. Diversification for better classification trees. Comput. Oper. Res., 33: 3185-3202.

[3] Webb, G. I., J. Boughton, Z. Wang, 2005. "Not So Naive Bayes: Aggregating One-Dependence Estimators". Machine Learning (Springer) 58 (1): 5–24.
[4] Kotsiantis, S.B., 2007. Supervised machine learning: A review of classification techniques. Informatica, 31: 249-268.
[5] Barros, R.C., M.P. Basgalupp, A.C.P.L.F. de Carvalho and A.A. Freitas, 2012. A survey of evolutionary algorithms for decision-tree induction. IEEE Trans. Syst. Man Cybern. Part C: Applic. Rev., 42: 291-312.
[6] Aitkenhead, M.J., 2008. A co-evolving decision tree classification method. Expert Syst. Applic., 34: 18-25.
[7] Kretowski, M. and M. Grzes, 2006. Mixed decision trees: An evolutionary approach. Proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery, September 4-8, Krakow, Poland, pp: 260-269.
[8] Gray, J.B. and G. Fan, 2008. Classification tree analysis using TARGET. Comput. Stat. Data Anal., 52: 1362-1372.
[9] Qin, B., Y. Xia, S. Prabhakar and Y. Tu, 2009. A rule-based classification algorithm for uncertain data. Proceedings of the 25th IEEE International Conference of Data Engineering, March 29-April 2, 2009, Shanghai, China, pp: 1633-1640.
[10] Giacometti, A., E.K. Miyaneh, P. Marcel and A. Soulet, 2008. A generic framework for rule-based classification. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, September 15-19, 2008, Antwerp, Belgium, pp: 37-54.
[11] Han, J., M. Kamber and J. Pei, 2011. Data Mining: Concepts and Techniques. 3rd Edn., Morgan Kaufmann Publishers, USA., ISBN-13: 9780123814791.
[12] Mozina, M., J. Demsar, M. Kattan and B. Zupan, 2004. Nomograms for visualization of naive Bayesian classifier. Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, September 20-24, 2004, Pisa, Italy, pp: 337-348.
[13] Vapnik, V., 1999. The Nature of Statistical Learning Theory, second edition. Springer Verlag. ISBN: 0387987800
[14] Bi, J. and T. Zhang, 2005. Support Vector Classification with Input Data Uncertainty. In: Advances in Neural Information Processing Systems, Volume 17, Saul, L.K., Y. Weiss and L. Bottou (Eds.). MIT Press, Cambridge, MA., USA., ISBN-13: 9780262195348, pp: 161-168.
[15] Dietterich, T.G., 2000. Ensemble methods in machine learning. Proceedings of the 1st International Workshop on Multiple Classifier Systems, June 21-23, 2000, Cagliari, Italy, pp: 1-15.
[16] Tan, P-N, M. SteinBach, and V. Kumar, 2005. Introduction to Data Mining, Addison Wesley. ISBN-10: 0321321367.
[17] Alashqur, A., 2010. RDB-MINER: A SQL-based algorithm for mining true relational databases. J. Software, 5: 998-1005.
[18] Alashqur, A., 2012. Using a lattice intension structure to facilitate user-guided association rule mining. Comput. Inform. Sci., 5: 11-21.

Author

Abdallah Alashqur is currently an associate professor at the Applied Science University in Amman, Jordan. Dr. Alashqur holds a Master's degree and a Ph.D. degree from the University of Florida. After obtaining his Ph.D. degree in 1989, he worked for around seventeen years in the USA (in industry) followed by nine years in Jordan (in academia). His research is mainly in the area of data mining and database systems.