Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

ANALYSIS OF MULTI-CLASS PATTERNS IN CORPORATE HOSPITALS USING TIME-VARYING DATABASES

ABHIMANYU PATRA

Ph.D Research Scholar, Department of Computer Science Engineering, Indira Gandhi Institute of Technology, Utkal University, Bhubaneswar, Odisha, India. Email: a_patra_2005@yahoo.com

SAROJANANDA MISHRA

Department of Computer Science Engineering and Application, Indira Gandhi Institute of Technology, Saranga, Odisha, India. Email: sarose.mishra@gmail.com

MANAS RANJAN SENAPATI

Department of Information and Technology, Veer Surendra Sai University of Technology, Burla, Sambalpur, Odisha, India. Email: manassena@gmail.com

RAJESH KUMAR BEHERA

Department of Mechanical Engineering, Krupajal Engineering College, Prasanti Vihar, Kausalya Ganga, Bhubaneswar, Odisha, India. Corresponding Author Email: rajesh_k_behera@yahoo.co.in

SUBHENDU KUMAR PANI

Department of Computer Science and Engineering, Krupajal Engineering College, Prasanti Vihar, Kausalya Ganga, Bhubaneswar, Odisha, India. Email: skpani.utkal@gmail.com

Abstract

Computerization has become a necessity in recent years. As a result, large amounts of digital data have accumulated across all industries. Data mining approaches have emerged to deal with this rich data and sparse information. Data mining is a process that uses computers to extract and explore hidden patterns in data. This knowledge discovery data mining process can be understood and utilized by users. Different weights are assigned to these attributes, depending on the value associated with each property. This study takes into account the weight prescribed by doctors. The first method presented is multi-class weighted association classification with confidence-based rule ranking, while the last step presented is genetics-based rule selection with a distributed multi-class weighted association classifier. Distributed weighted association classification is used as a consequence of the results, which reduces the cost of communication, while maintaining the advantages of the centralized approach. The % accuracy of multi-class classification improves when weighted associative classification is used.

Index Term: Data Mining, Artificial Neural Network, Fuzzy Logic, Time series, MCWAC, Pruning Rule, and MCWACGRS

1) INTRODUCTION

Data mining is a method for discovering and extracting hidden patterns with the use of computers. Many data mining approaches can be used to extract knowledge from the datasets. The data mining techniques help in efficiently uncovering useful patterns and information from massive amounts of data [1].

The vast availability of enormous volumes of actual data and the impending necessity to transform such real data into valuable information are the main factors drawing a lot of attention to data mining. Predictive analysis (supervised) and descriptive analysis

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

(unsupervised) are said to be the two main kinds of data mining [2-4]. In the prediction model, all the attributes do not have the same importance in predicting the class label.

Algorithms in predictive analysis make it easier to find patterns and deduce models that aid in the forecasting of future data classes. According to Feng et al. and Han et al. [5, 6], classification, prediction, regression, and time series analysis methods fall within the category of predictive analysis.

Experts in the field can make decisions more effectively with the help of predictive analysis. Nowadays, predictive analysis has also adopted descriptive analysis. This is where the suggested research project is concentrated.

In this paper, multiclass classifiers based on weighted associative classification algorithms are proposed [10-11]. Association rule discovery techniques have been successfully applied in recent years to create precise classifiers [12-16]). The majority of associative classifiers focused on creating two class classifiers and took into account unit weight for every characteristic.

Designing multiclass classifiers with attribute value-based weights is the main goal of the proposed study. For the purpose of creating rules, the physicians' allocated weights are taken into consideration. We offer three methods based on the multiclass weighted associative classification algorithm.

A statistical parameter-based rule pruning method called confidence-based multiclass weighted associative classification is initially proposed. A genetic rule pruning procedure based on an external population technique has then been utilized to increase the accuracy of the final multiclass classifier. The binary length representation for each attribute in this genetically based technique is altered for the beginning population (external population), crossover, and mutation as well.

Ultimately, a distributed environment has been added to the multiclass weighted associative classification method to make it work in real-time scenarios. The importance of weights related to the item set values in the market-basket analysis served as the impetus for their proposal of multiclass weighted associative categorization [17]. The translation of a dataset into its equivalent weighted dataset is the first step in the proposed weighted associative classification method.

Rule generation and rule trimming are the two stages of the proposed multiclass weighted associative classifier. Using multi-class, multi-label associative classification, a full set of Class Association Rules (CAR) is identified and constructed during the rule creation phase. Multiclass Multilabel Based Associative Classification (MMAC), which identifies classes based on association rule mining in a single database scan [18]. This makes the rule generating process less time-consuming.

The MMAC is expanded for weighted association rule mining in the suggested ways. All of the weighted attribute combinations are regarded as rules in this phase. Hence, volumes of rules generated are huge. Then comes the role of the second phasenamely, rule pruning. Identifying the effective set of rules is a challenging task. In the first

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

technique, user defined interestingness measures such as support and confidence are utilized along with rule ranking for classifierconstruction.

The formulae for calculating support and confidence are shown in Eq. 1 and 2, respectively.

$$Support\ (x \Rightarrow y) = \Sigma_{d \in D} \frac{\mu_{xy}(d)}{|D"|} \quad (1)$$

$$Confidence\ (x \Rightarrow y) = \Sigma_{d \in D} \frac{\mu_{xy}(d)}{\mu_x(d)} \quad (2)$$

Where, x = Antecedent part of the rule

y = Consequent part of the rule is class label d

$\mu_{xy}(d)$ = Weight associated to the rule $x \Rightarrow y$ for class label d

$|D"|$ = Total number of records

The choice of efficient rules has the most significant impact on the classifier's effectiveness. Instead of choosing the best rules for classifier model generation only based on confidence, the suggested technique uses a Genetic-Algorithm (GA) to identify the best rules for classifier construction, with accuracy serving as the goal function. In Eq. 3, the accuracy formula is presented.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (3)$$

Where, TP: True Positive is the quantity of positive cases that have been accurately identified as being in the positive class. TN: True Negative is the quantity of negative cases that were accurately categorized as being in the negative class. FP: False Positive (FP) refers to the number of negative cases that were mistakenly placed in the positive class. FN: False Negative (FN) refers to the number of positive cases that were incorrectly categorized as negative.

The findings of the experiments show that choosing a small set of rules with good classification accuracy is facilitated by genetic-based rule selection. The expansion of the multiclass weighted associative classification method in the distributed environment is the final effort for the distributed real-time dataset.

The study reveals the weighted dataset is employed as an input for all the three proposed techniques based on the multiclass associative classification. The proposed methods are successfully tested on two distinct heart datasets, and the outcomes are shown at the conclusion.

2) ADOPTED METHODOLOGY

2.1 Multiclass weighted associative classification (MCWAC)

The advantages of multiclass multi label over other associative classification techniques are as follows. It creates classifiers with rules with many labels, using a powerful technique for finding the rules that only needs to scan the training set once, and

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

combines frequent item set discovery and rules generation in one step to save space and speed up processing. All attributes in this algorithm only have a single unit of weight. Yet, the importance of assigning values that differ has not been addressed.

As a result, the objective behind the suggested strategy is to elevate attribute values. As a result, for CAR generation, all of the proposed methods in this research use multiclass multi label based associative generation.

In the proposed system, the foremost step is the transformation of a dataset into a weighted dataset; then the multiclass algorithm based frequent item set discovery for rule generation is employed.

The proposed multiclass weighted associative classification has three steps. A dataset is converted into a weighted dataset as the first step. As it is a multi-class classifier, the second stage of the technique is rule generation, where frequent item sets are found and association rules are constructed for every combination of class labels.

The third stage, where the enormous number of rules generated in the second stage is trimmed, is crucial for the classifier model design.

Rule ranking-based rule trimming is proposed in the multiclass weighted associative classification model. Based on the rule ranking method, important rules are found. Here, by using confidence as the only criterion for interestingness in the suggested system, unnecessary rules are removed.

The proposed approach focuses on creating a multiclass classifier; a support measure limits rule generation for a small number of classes that lack sufficient training data. Support demonstrates the frequency of the rule item set across the full dataset. When support is taken into account as an interestingness criterion for rule selection, there is a potential that one or two class-related rule sets will be lost if there are few training records that correspond to the classes.

Fig. 1 shows a diagrammatic representation of the stages involved in this technique. Data transformation, frequent rule item set generation, association rule production for all classes, and lastly rule ranking and rule trimming for the creation of the final classifier model are the three stages of the proposed system.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

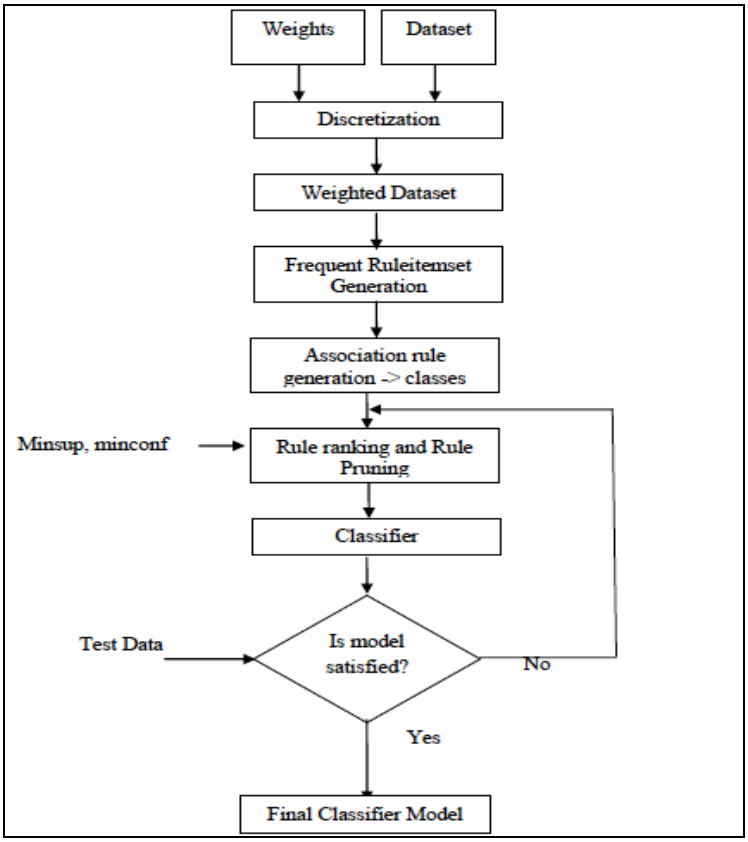

Fig 1: Proposed multiclass weighted associative classification flow chart

2.1.1 Transformation of Data

Data-transformation is the process of converting a dataset into a weighted dataset. The first step is to discretize the attributes and give the discretized range the proper weights. According to the advice of the doctors (domain experts), discretization and weights to the attributes are assigned in the planned work. In this study, two heart data sets are taken into account when building a multiclass classification system. The two heart datasets that were used in the experiment. Both datasets have a total of fourteen attributes, including the class label attribute. Including the six examples with missing attribute values, the UCI heart dataset contains 300 instances. In the private dataset, 210 cases have been gathered. The class distribution is shown in Table 1 and the weights that the doctors allocated to each attribute range are shown in Table 2. Throughout the course of our research, these values have not changed.

Table 1: Class Distributions

Data base	Class-0	Class-1	Class-2	Class-3	Class-4	Total
Private heart dataset-KIIMs hospital	161	51	36	35	13	300
Private dataset- Kalinga hospital	107	49	24	18	12	210

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

The weights assigned by the physicians for each attribute range and weights for various discretized range are given in Table 2.

Table 2: Attribute Weights and Range

Attribute Number	Attribute Name	Attribute Description	Range	Weight
1.	Age	Age in years	<40 41-60 >60	0.5 0.8 0.9
2.	Sex	1= male; 0= female	1 0	0.9 0.1
3.	Cp	Kind of chest pain Value 1: usual angina pain Value 2: unusual angina pain Value 3: non-anginal pain Value 4: asymptomatic pain	 1 2 3 4	 0.9 0.6 0.3 0.1
4.	Trestbps	Resting blood pressure on admission to the hospital (in mm Hg)	>140 121=140 <120	0.9 0.5 0.1
5.	Chol	Serum cholesterol (in mg/dl)	<200 >200	0.1 0.9
6.	Fbs	fasting blood sugar (> 120 mg/dl) 1=true 0=false	 >120 <120	 0.9 0.1
7.	Restecg	Resting electro-cardiographic results *Value 0*: normal *Value 1*: having ST-T wave abnormality (T- wave inversions and / or ST elevation or depression of > 0.05 mV) *Value 2*: showing probable or definite left ventricular hypertrophy	 0 1 2	 0.1 0.5 0.9
8.	Thalach	Maximum heart-rate achieved	>100 >120	0.5 0.9
9.	Exang	Exercise induced angina 1 = Yes 0 = No	 0 1	 0.1 0.9
10.	Slope	The slope of the peak exercise ST-segment Value 1: up sloping Value 2: down sloping	 1 2	 0.9 0.1
11.	Ca	Number of major vessels (0-3) colored by fluoroscopy	0 1 2 3	0.1 0.25 0.5 0.9
12.	Num	Diagnosis of heart diseases (angiographic disease status) Value 0: < 50% diameter narrowing Value 1: > 50% diameter narrowing	0 1 2 3 & 4	No-risk Low-risk Medium-risk High-risk &Very-high risk

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

One of the attributes, for instance, is age, which is discretized into three categories. The first range is less than forty (low), the second range is forty one to sixty, and the third range is greater than sixty (high), with the corresponding values being 0.5, 0.8, and 0.9. Several types of discretization occur depending on the application domain because automatic discretization is not always effective. With the generalized model, classification created for a particular application will not function well. This study focuses on creating multiclass classifications for estimating different heart disease risk levels. According to the experts' (physicians') guidance, the proposed multiclass weighted associative classification technique discretizes qualities into two, three, or four ranges. For each discretized range, subsequent relevant weights have been assigned. Lastly, a dataset is converted to a weighted dataset by substituting the weight of its associated attribute. From this point forward, weighted dataset is used as an input for subsequent processing.

Ages under forty are considered young in the discretization. As a result, the young category's weight is 0.5. Now, 0.5 is used in place of 40 in the data point. Similar to how 1 represents a female patient, 0.9 is substituted for 1 to represent a male patient's weight, and the cholesterol value is discretized into low risk and high risk depending on whether it is less than 200 or more than 200. Low risk cholesterol is given a weight of 0.1, whereas high risk cholesterol is given a weight of 0.9.

2.1.2 Rule Generation of Multiclass Weighted Association

The second step of the proposed system uses the MMAC technique from Thabtah et al. (2004) to improve the effectiveness of frequent items discovery and rules development. Multiclass classification has been accomplished using the weighted multiclass associative classification extension of the MMAC approach. The proposed technique's rule generation mechanism counts the occurrences of individual items in the training data once, then selects those that pass the minsupport and minconfidence requirements and saves them in rapid access data structures with their occurrences (rowIds). By simply intersecting the rowIds of the previously discovered frequent single items, it is then possible to easily identify the remaining likely frequent items that involve several characteristics. For rules involving several items, the rowIds for frequently occurring single items are useful information that can be utilised to quickly locate objects in the training data.

Consider the often occurring single items A and B to better understand the concept of this method. If the rowIds sets of these two items are intersected, the resulting set should reflect tuples in which A and B occur together in the training data. In order to determine whether or not A and B are frequent itemsets and may be included or not in a candidate rule of the classifier, it is therefore simple to discover the classes linked with A and B in which the support and confidence can be evaluated and calculated. This method is very effective in lowering space and time complexity because the training data was only searched once to find and generate the rules. Rules can be removed after all patterns have been generated. This phase's goal is to create every feasible set of rule item sets across all classes. There are a lot of regulations produced.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

2.1.3 Ranking and Pruning Rule

Due to the size of the total number of rules generated, interesting and practical rules are filtered using support and confidence metrics. The third phase of the suggested technique is responsible for this task. The proposed approach focuses on creating multiclass classifiers; the support measure limits rule generation for a small number of classes that lack sufficient training data. Support demonstrates the frequency of the rule item set across the full dataset. When support is taken into account as the interestingness criterion for rule selection, there are occasionally chances of losing the rule item set for one or two classes where there aren't many training records for those classes. Thus, the support threshold has received less attention in this research endeavor. Confidence measure exposes the presence and significance of the rule item sets in specific classes. Therefore, all the techniques proposed in this and the following chapter adopt confidence based pruning.

For each rule produced in accordance with Eq. 1 and 2, support and confidence values have been determined. The rule item is retrieved with confidence and support values that are higher than the minimum confidence and support values, or minsupport, respectively. User defined interestingness measures include minsupport and minconfidence. The suggested method determines whether or not an itemset passes the minconfidence level after identifying it as a frequent itemset. If the itemset (rule) meets the minimum confidence criterion, it is put in the classifier's candidate rule list. If not, the ruleitem will be thrown away. The classifier refers to all rule items that are still valid with a minimum confidence level of above as candidate rules.

Rules that don't receive at least the minimum amount of support and trust are discarded. The confidence rating of the remaining rules determines their order. With a two class classifier, redundant rules are typically identified as being ineffective and removed. Less rules are created in a few classes when using a multiclass classifier. As a result, when creating a multiclass classifier, combinations of survived rules for each class are taken into account. Based on the precise classification of test data, the classifier model's accuracy is evaluated. The effectiveness of the suggested strategy is described in the following subsection.

The public heart dataset is used to evaluate the proposed Multi Class Weighted Associative Classification (MCWAC), and the results are given in Table 3. The number of rules produced by the suggested multiclass weighted associative classification is shown in the table clearly. The second row of Table 3 shows that a significant number of Weighted Multi Class Multi Label Associative Classification (WMCMLAC) rules have been created. It causes the production of overfit classifier models for varied risk levels. Rule pruning is therefore necessary. Pruning is done in this research project based on the confidence parameter. A truly intriguing measure for gauging interestingness is confidence, which demonstrates the real connection between the antecedent rules and the consequent class label. The number of rules generated for various support and confidence values are displayed in a chart view in Fig. 2. Support assesses the frequency of the rule in the full data set, hence it is useless for creating multiclass classifier models.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

Table 3: Results of the Proposed Weighted Associative Classification with Confidence Based Rule Pruning for Public Heart Dataset with Varied Support and Confidence Thresholds

Number of Rules	Support 2%	Support 5%	Support 10%	Support 20%	Support 30%	Support 50%
WeightedRecords	300	300	300	300	300	300
WMCMLAC	93950	69690	38925	19310	13985	8385
Pruned Rules(confidence >15%)	1043	640	396	209	145	77
Pruned Rules(confidence >25%)	929	589	365	191	134	73
Pruned Rules(confidence >50%)	523	232	141	114	105	71

The frequency of the rule for the particular class is required for the proposed work. The confidence parameter has therefore been given prominence. Table 4 shows that an increase in the support value results in zero rules at risk level four. High confidence rules (>50%) are retained for the final classifier model due to the relevance of confidence.

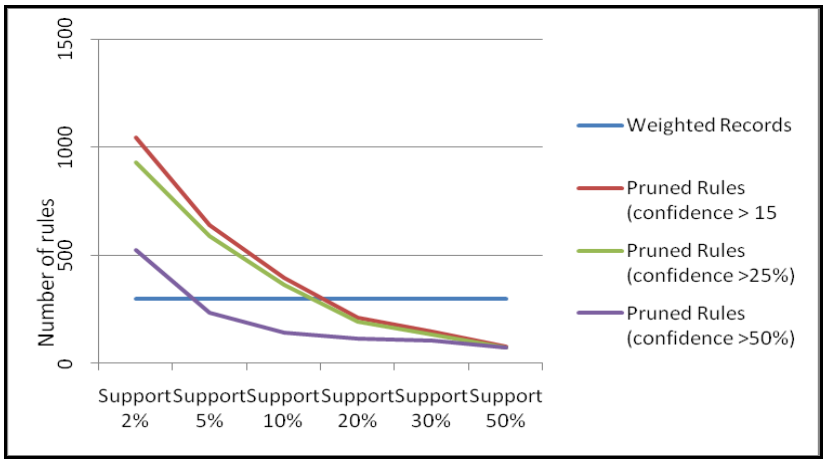

Fig 2: Chart view of the result of the proposed weighted associative classification with confidence based rule pruning for public heart dataset with varied support and confidence thresholds

The rule distribution for various risk levels is clearly shown in Table 4. For the risk 0 and risk 1 classes, it is acknowledged that a sufficient number of rules are available to build a classifier model. Fewer rules apply in Risk 2. There are very few criteria for the risk 3 and risk 4 classes, and none have been chosen for the support value of 50% for risk 2, 3, or 4. There are undoubtedly too few rules present to generate a multiclass classifier model. There is room for improvement in the suggested method.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

Table 4: Results of the Proposed Multiclass Weighted Associative Classifier for Various Risk Level

		Risk Level	Support 2%	Support 5%	Support 10%	Support 20%	Support 30%	Support 50%
No. of Rules (Weighted Associative Classification) Confidence >50	KIIMs Dataset	0	474	188	103	96	90	65
		1	24	22	20	11	9	06
		2	17	14	10	03	03	--
		3	05	05	05	04	03	--
		4	03	03	03	--	--	--
No. of Rules (Weighted Associative Classification) Confidence >50	Kalinga Dataset	0	438	219	177	126	81	71
		1	49	44	38	23	19	13
		2	27	21	21	18	12	03
		3	12	09	07	07	06	01
		4	05	04	04	03	--	--

The multiclass classifier model has room for improvement, according to experimental results. Building a multiclass classifier requires more than just the confidence parameter. Another parameter used to assess any classifier is accuracy. Hence, an evolutionary method is used with accuracy as the objective function to increase the multiclass classification accuracy. The concept of genetically based rule selection is developed in the section that follows.

2.2 Proposed Multiclass Weighted Associative Classification with Genetic Based Rule Selection (MCWACGRS)

The level of interestingness of the rules produced by association rule-mining is constrained by support and confidence. The quantity of rules produced using support and confidence is insufficient for building multiclass classifier models. Accuracy is a crucial measurement for rules creation. The number of records that were correctly categorized using a rule is shown by accuracy. It is suggested to use multiclass weighted associative classification with rule selection based on genetic algorithms with accuracy as the objective function. The creation of weighted datasets is the initial step. The second stage of the suggested method is rule creation, and it is identical to the method suggested in the last section, in which frequent item sets are found and the association rules are constructed using the MCMLWAC methodology for every combination of class labels. A large number of rules that were developed in the first step are prepruned with support and confidence thresholds in the third stage, which is rule pruning. Preliminary rule pruning (Prepruning) has been performed by fixing a lower threshold value for support and a higher value for confidence, despite not fully understanding the importance of support in multiclass classification. The genetic algorithm for rule selection that uses an external population technique as its last step has accuracy set as one of its objective functions. Eq. 3 provides the formula for computing the support. The external population-based genetic algorithm is used to develop a multiclass classifier by helping to determine the precise rules. Prepruned rules that were extracted in the second stage are used as the genetic approach's initial population. The suggested technique uses the external population-based GA technique,

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

which is described in the following subsections. In Fig. 3, the steps of the suggested technique are represented diagrammatically.

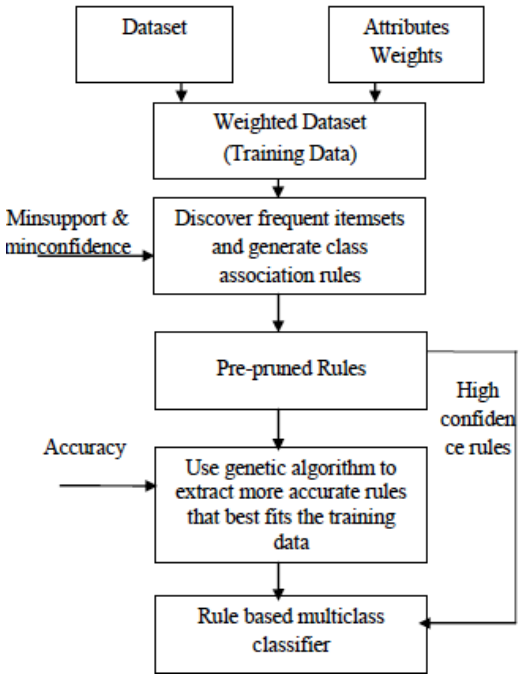

Fig 3: Flow diagram of the proposed multiclass weighted associative classification with genetic based rule selection

2.2.1 Transformation of Data

According to the previous subsection's explanation, the dataset is transformed into a weighted dataset. Several ranges of discretization occur depending on the application domain. The external population-based genetic algorithm (GA) is encouraged for rule selection by this variable discretization. According to the doctors' recommendations, the qualities in the proposed system are discretized into two, three, four, or five ranges. Every discretized range is then given a new set of weights. A dataset is finally replaced with the weights assigned to its relevant attributes, transforming it into a weighted dataset.

2.2.2 Rule Generation and Rule Prepruning

The training data is once again scanned to find and produce the rules. It is possible to generate rules from all conceivable rule itemsets in the frequent itemsets once all the patterns have been generated. This is comparable to the rule generation phase of the earlier confidence based weighted multiclass classifier. The rules with confidence and support values higher than the minimum values are extracted. Prepruned rules are provided as a selection input for genetically based rules.

2.2.3 Rule of Genetic Based Selection

The number of rules produced by the confidence-based weighted associative classification that was previously proposed is insufficient for building the multiclass

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

classifier. Thus, the genetic algorithm's evolutionary optimization technique is used to extract precise rules from the training dataset. In order to address generic optimization issues with huge search areas, genetic algorithms (GAs) were created. The stages taken in the evolutionary computation of the genetic algorithm are described in Fig. 4.

Input : Prepruned rules, Threshold for accuracy
Output : Accurate rules

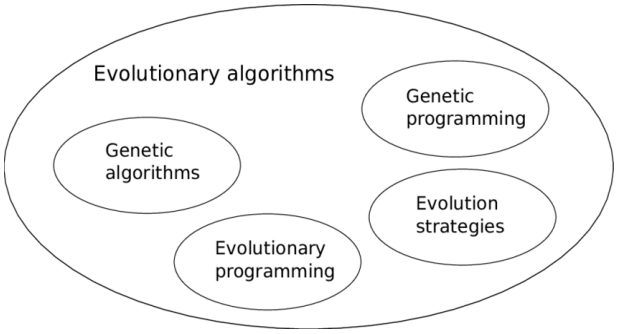

Fig. 4. Evolutionary Computation.

The population's genetic diversity will be increased by the mutation operator. Solutions that are encoded are immediately subjected to the mutation operators. As a result, the representation specified for the solutions must be used to construct the mutation.

2.2.4 Results and Analysis

The public heart dataset from the UCI repository serves as the basis for the investigation of the suggested genetic-based selection. This public dataset's nature is incredibly hazy. In order to generate rule items often, metrics support and confidence play a crucial role. The weighted dataset that has been transformed serves as the technique's input. The weighted average of the attribute values provided by the doctors is shown in Table 2 and is used to replace the attributes in the record set. The technique suggested in the previous section uses a similar second stage of rule development. The weighted dataset's rules are produced using MCMLAC. For each and every class, all conceivable rules are constructed. As a result, there are a lot of produced rules. The second row of Table 5 displays the total number of rules produced for each support value. The support and confidence for each rule in each class are concurrently determined in this MCMLAC technique along with the rule generation. Hence, compared to other apriori-based techniques, the time complexity is lower. All of the MCMLAC with support and confidence values are produced by a single scan of a dataset. Prerunning in the third stage involves setting 2% as the minimum support and 75% as the minimum confidence. The results are shown in Tables 3 and 4. The next stage is genetic based accurate rule selection. The genetic algorithm uses the rules that were previously trimmed in this stage. That is, the earlier stage's output is used as the stage's input. In the second stage, the encoded rules are trimmed from the initial population. The processes are repeated 500 times after crossing and mutation. The true positive (TP), true negative (TN), false positive (FP), and false negative values for the genetically based created rules are kept in a file (FN). There are a large number of rules

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

produced overall for different danger levels. The number of rules generated at each level, both with and without optimization, is shown in Table 5. The number of rules that have been pruned with accuracy after pruning is shown as > 80% in the fourth row of the table. It has been noted that the number of genetic rules with accuracy >80% for risk level 0 (4, 3, 2, 1, 1, 0) is extremely low (307, 284, 217, 187, 174, 124). The issue is because there are a lot of false positive and false negative rules generated for the risk level 0 classes, and these rules are also present in many other risk level rule sets. Hence, a multiclass classifier cannot be constructed using only genetically based results. Consequently, the suggested multiclass weighted associative classification may be strengthened by combining the confidence-based rule pruning technique and the genetic-based rule selection technique. So, in this chapter, a multiclass weighted associative classifier is built with genetically based selected rules for other risk levels and high confidence rules for risk level 0. As a result, the suggested multiclass weighted associative classifier with genetically based rule selection for early heart disease prediction is successfully created. The accuracy of the final classifier is also tested using both private and public datasets. Java is used to implement each technique that has been suggested. The suggested classifier's accuracy is compared to that of the RIPPER and PART decision tree classifiers currently in use. Table 6 depicts the accuracy comparison between the existing techniques and the proposed multiclass weighted associative classification with genetic based rule selection technique against both private heart dataset.

Table 5: Results of the Proposed Multiclass Weighted Associative Classification with Genetic Algorithm Based Rule Selection (MCWACGRS)

KIIMs Dataset	Risk Level	Support 2%	Support 5%	Support 10%	Support 20%	Support 30%	Support 50%
Rules generated WMCMLAC		93950	69690	38925	19310	13985	8385
Total number of rules generated using GA	All	929	864	739	556	535	331
	0	307	284	217	187	174	124
	1	201	199	175	122	118	71
	2	153	137	119	83	86	44
	3	149	138	130	94	90	50
	4	119	106	98	70	67	42
Number ofrules >80%	All	468	412	365	262	259	136
No. of Rules using MCWACGRS with Accuracy >80%	0	4	3	2	1	1	0
	1	156	133	110	73	85	42
	2	109	94	83	61	60	29
	3	102	99	92	71	63	34
	4	97	83	78	56	50	31

WEKA (a well-known data mining tool) was used for the experimental study of PART and RIPPER, and the results are presented in Table 6 for comparison. These two conventional classifiers are both classifiers. Java is used to implement the accuracy of the existing MMAC, and Table 6 also shows the accuracy attained using the current method. All attribute values in MMAC were solely mapped to integer values. The weights allocated to physicians in this study are ambiguous. As a result, the proposed weighted multiclass classifier's accuracy has improved. Traditional classifiers such

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

decision tree-based PART, RIPPER, etc. focus exclusively on two class classifiers in comparison to the existing approaches. Without taking into account the attribute weights, a multi class multi label based associative classification created a multi class classifier. The public dataset's nature is incredibly hazy. As a result, the private dataset's accuracy is higher than the public dataset's. The private heart dataset's attribute values-based weight assignment and class assignment are carried out with the assistance of the same group of subject-matter experts (physicists). As a result, the private dataset's accuracy is greater than the public cardiac dataset's.

Table 6: Comparison of Accuracies of the Proposed MCWACGRS Algorithm and the Existing Algorithms

Dataset	Existing Algorithms			Proposed Algorithm
	PART (Decision tree)	RIPPER	MCMLAC	MCWACGRS
Classifier category	Two	Two	Multiclass	Multiclass
Public dataset Accuracy in %	67	71	75	84
Private dataset Accuracy in %	73	76	78	88

The proposed multiclass weighted associative classifier is extended for distributed environments, as the majority of the dataset are distributed in nature, in order to take advantage of the advantages of the proposed multiclass weighted associative classifier with genetic based rule selection in real life scenarios. As a result, a new distributed multiclass weighted associative classification is suggested, and it is further explained in the part that follows. There are two ways to apply data mining techniques on geographically dispersed data sources: centralized model and distributed model. In a centralized model, the necessary data from numerous sources is collected at one location, and a mining technique is applied to the combined data. It yields precise results but takes a lot of time and money for communication. Given the sensitivity of the dataset, security must be provided for data transit. Whereas in a distributed setting, local sites will do partial mining and the results from those locations will be combined at a single site. When the same condition and threshold are specified in the centralized model, the accuracy of distributed models will be reduced. Distributed models simplify algorithms and lower communication costs. A distributed associative classification algorithm to reduce time complexity and used a function [(item-1) % N] for distributing items to N number of processors. With minimum communication between processors, locally generated subset ofclassification rules are finally combined to make a classifier. In their work, the dataset need not be partitioned. It further reduces the communication cost. Ultimate objective for any parallel or distributed associative classification is to reduce communication cost and less inters process communication designed a parallel model for CBA system [19-24]. The multiclass weighted associative classification introduced in the preceding part is expanded for distributed environments in this study because there is a dearth of research on the subject. The results are reported in the next subsection.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

2.2.5 Design of Proposed System

By successfully utilizing the knowledge regarding the data gathered at numerous sites, the suggested system aims to create a multiclass classifier model. It improves the model even more by including the typical associated patterns linked to the dataset across different locations. In recent years, distributed databases where data is stored in various shared nothing machines connected through internet, grid or cloud. The suggested approach is similar to distributed data mining in that it distributes data across a number of sources horizontally.

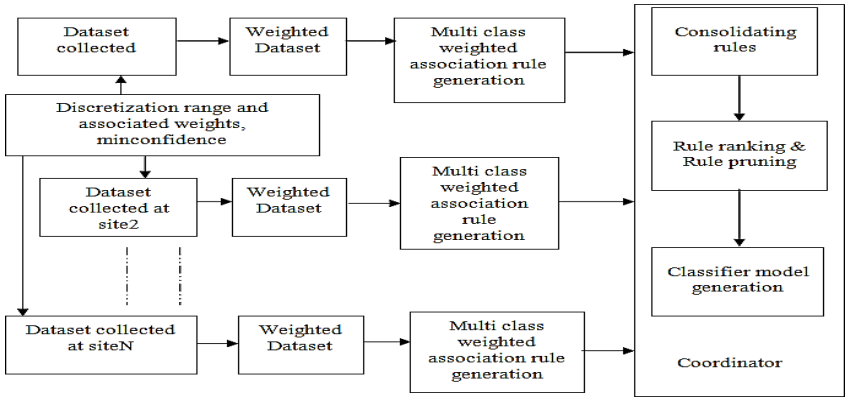

Fig 5: Design diagram of the proposed distributed multiclassweighted associative classification system

The proposed system modifies [25-29] few factors of the distributed shared nothing environment. The first is to incorporate the advantages of a centralised model into a distributed model; the task of building a classifier model is taken on by a common coordinator. Data gathered at each site is used as a training dataset in the centralized model, and it is sent to the coordinator for consolidation and the creation of classifier models. Semi-pruned rules and their confidence values are sent to the coordinator in the proposed system in order to lower communication costs and lower the security risk associated with delivering the dataset in its current state. The rules produced using the training data dispersed across many sources is used as a whole in the distributive environment to create classification models. The suggested distributed system's architecture diagram is given in Fig. 5.

2.2.6 Experiments, Results and Analysis

Starting the investigation is the public heart dataset. The public (UCI) heart dataset is separated into three datasets for analysis [30-34]. Table 7 lists the preliminary and final outcomes at each location. More rules are discovered for classes with greater risk levels when the minconfidence threshold is lowered to 40%. The only threshold taken into account for this distributed model is minconfidence. Table 7 lists the total number of rules produced at each stage and for different risk levels. It also demonstrates that for higher risk levels, a manageable amount of rules are developed. The workload on the coordinator site is decreased by the proposed distributed technique because partial mining has been finished at each location.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

Table 7: Results of the Proposed Distributed Weighted Associative Classification

Site	Number of Records	Total No. of. Rules MCMLWAC	Number of Rules	
			Confidence >40	Risk levels 0,1,2,3,4
Site 1	99	43100	415	341
				31
				16
				15
				12
Site 2	99	39105	380	308
				28
				14
				18
				12
Site 3	99	33480	318	245
				29
				20
				13
				11

While the majority of past research studies concentrated on two-class classifiers, the proposed multiclass classifier developed in this research effort aids in the early detection of disorders.

3) CONCLUSION

The certain attributes are given varying weights based on the value they are linked with. Weights assigned by the doctors are taken into consideration in this investigation. The first method discussed is multiclass weighted associative classification with confidence based rule ranking. A genetically based rule selection mechanism is introduced in the next enhancement and presented as a final step is a distributed multiclass weighted associative classifier. The proposed techniques include three improvements. Firstly, the weights of the attributes are not uniformly discretized and each attribute holds different range of weights according to the significance. Secondly, the number of scans required for association rule generation is only one due to the adoption of multi class multi label association rule generation algorithm and finally, evolutionary computation with accuracy as the objective function is realized for rule selection.

The following conclusions are drawn:

1. The distributed weighted associative classification is suggested in order to lower communication costs while maintaining the benefit of the centralized approach.
2. Heart datasets were used for testing all three of the proposed approaches, and the results highlight their potency.
3. Weighted associative classification boosts the multiclass classification's accuracy %.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

Declaration of interests:

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval:

This article does not contain any studies with human participants or animals performed by any of the authors.

References

1) A.A. Aljumah, M.G. Ahamad. M. K. Siddiqui, Predictive Analysis on Hypertension Treatment using Data Mining approach in Saudi Arabia, Intell. Inf. Manag. 3 (2011) 252-261.

2) C.S. Dangare, S.S. Apte, Improved study of heart disease prediction system using Data Mining Classification techniques, Int. J. Comput. Appl., 47(10) (2012) 44-48.

3) L. Duan, W. N. Street, E. Xu, Data Mining methods in the creation of a clinical recommender system & Enterprise Information Systems, Health Inf. Sci. Syst. 5(2) (2011)169-181.

4) F. Tao, F. Murtagh, M. Farid, Weighted Association Rule Mining using Weighted Support and Significance Framework Proceedings of the ninth ACM SIGKDD, Int. Conf. on Know Discov. Data Min. 2 (2003) 661-666.

5) J. Han, J. Pei, Y. Yin, R. Mao, Mining frequent Patterns without candidate generation: A frequent pattern tree approach, Data. Min. Knowl. Discov. 8(1) (2004) 53-87.

6) A. Kusiak, Feature Transformation Methods in Data Mining, IEEE Trans. Compon. Packag. Manuf., 24(3) (2001) 214-221.

7) S. Gupta, D. Kumar, A. Sharma, Data Mining Classification Techniques applied for breast cancer diagnosis and prognosis. Indian J. Comput. Sci. Eng. 2(2) (2011) 188-195.

8) E. Mansoori, M. Zolghadri, S. Katebi, A steady-state genetic algorithm for extracting fuzzy classification rules from data, IEEE Trans. Fuzzy. Syst. 16(4) (2008) 1061-1071.

9) K. Srinivas, B. Kavihta Rani, A. Govrdhan, Applications of Data Mining techniques in Healthcare and Prediction of heart attacks, Int. j. comput. sci. eng. 2(2) (2010) 250-255.

10) I. Pramudiono, M. Kitsuregawa, Shared nothing parallel execution of FP-growth, Adv. Knowl. Discov. Data Min. 2637 (2003) 467-473.

11) S. Gupta, D. Kumar, A. Sharma, Data Mining Classification Techniques applied for breast cancer diagnosis and prognosis, Indian J. Comput. Sci. Eng, 2 (2011) 188-195.

12) B. Liu, W. Hsu, Mun Lai-Fun., H. Lee, Finding Interesting Patterns Using User Expectations, IEEE Trans Knowl Data Eng., 11(6) (1999) 817-832.

13) B. Liu, S. Y. Philip, Top 10 algorithms in data mining, Knowl Inf Syst. 14 (2008) 1-37.

14) F. Thabtah, A review of associative Classification Mining, Knowl Eng Rev. 22(1) (2007) 37-65.

15) S. Lu, H. Hu, F. Li, Mining weighted association rules, Intell. Data Anal. 5(3) (2005) 211-225.

16) F. Thabtah, P. Cowling, Y. Peng, MMAC: A new multi-class, multi-label associative classification approach, in proceedings of the 4th IEEE International Conference on Data Mining, Brighton, UK, (2004) 217-224.

17) D. Mokeddem, H. Belbachir, a Distributed Associative Classification Algorithm, Intel. distrib. comput. Springer, (2010) 109-118.

ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/4CBSX

18) R. Agrawal, J. C. Shafer, Parallel Mining of Association Rules, IEEE Trans Knowl Data Eng. 8(6) (1996) 962-969.

19) M. Zaki, Parallel and Distributed Data Mining: An Introduction, in Large-Scale Parallel Data Mining, IEEE Trans Knowl Data Eng. 5 (2000) 1-23.

20) B. RaghuRam, Gyani, B. Hanmanthu, Fuzzy Associative Classifier for Distributed Mining, Int. J. Comput. Appl. 9 (2012) 1-5.

21) G. Thakur, C. J. Ramesh, A Framework for Fast Classification Algorithms, Int. J. Inf. Theories Appl. 15 (2008) 363-369.

22) K. Aftarczuk, Evaluation of selected data mining algorithms implemented in Medical Decision support systems', Thesis no: MSE-2007-21, September 2007, School of Engineering, Blekinge Institute of Technology, Sweden, (2007).

23) S. Patil, B. Kumaraswamy, Intelligent and effective heart attack prediction system using Data Mining and Artificial Neural Network, Eur. J. Sci. Res. 31(4) (2009) 642-656.

24) D. Swapnil, W. Avinash, Guidelines of Data Mining Techniques in Healthcare Applications, Int. j. adv. res. comput. 2(4) (2013) 1393-1397.

25) M. R. Senapati, S. P. Das, P. K. Champati, P. K. Routray, Local linear radial basis function neural networks for classification of breast cancer data. In: PISER 16, 02 (2014) 033–042.

26) A. Bahrammizaee, A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert systems and hybrid intelligent system. Neural Comput Appl. 19(8) (2010)1165–1195.

27) J. M. Tomczak, A. Gonczarek, Decision rules extraction from data stream in the presence of changing context for diabetes treatment. Knowl Inf Syst. 34(3) (2013) 521–546,.

28) K. Thangavel, P. P. Jaganathan, P. O. Easmi, Data Mining Approach to Cervical Cancer Patients Analysis Using Clustering Technique, Asian J. Inf. Technol. 5(4) (2006) 413-417.

29) M. G. Tsipouras, T. P. Exarchos, D. I. Fotiadis, A. P. Kotsia, K. V. Vakalis, K. K. Naka, L. K. Michalis, Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling, IEEE trans. inf. technol. biomed. 5 (2008) 447-458.

30) I. Ullah, Data Mining algorithms and Medical Sciences, Int. j. comput. sci. inf. technol. 2(6) (2010) 127-136.

31) Y. Unil, A new framework for detecting weighted sequential patterns in large sequence databases, Knowledge based systems, Asian J. Inf. Technol. 21(2) (2008) 110-122.

32) W. Wang, J. Yang, P. Yu, Efficient mining of weighted Association rules (WAR), Knowl Discov Data Min.14 (2012) 270-274.

33) S. K. Wasan, V. Bhatnagar, H. Kaur, The impact of Data Mining techniques on Medical Diagnostics, Data Sc. Jour. 5 (2006) 119-126.

34) C. Zuoliang, C. Guoqing, Building an Associative Classifier based on Fuzzy Association Rules, Int. J. Comput. Intell. Syst., 1(3) (2008) 262-273.