Implementing Hybrid GA & Fuzzy KNN to Optimize the Classification of Fuzzy KNN

Simarpreet Kaur¹, Jyoti Arora²

¹Student, Desh Bhagat University
²Assistant Professor, Desh Bhagat University

Abstract: In this paper data mining is the process of fetching hidden knowledge from a wide store of raw data. The knowledge must be new, and one must be able to use it. CRM is an information industry term that helps an enterprise to manage customer relationships in an organized way and helping the company to provide better services to its customers. In this paper to overcome the problems of CRM database a new hybrid algorithm is introduced which will be the combination of GA and Fuzzy KNN classification.

Keywords: data mining, Fuzzy KNN, Genetic Algorithm, Precision, Recall, f-measures, g-mean, sensitivity & specificity

1. Introduction

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

This white paper provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today’s business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

1.1 Customer Relationship Management (CRM)

CRM is the core business strategy that integrates internal processes and functions of the organization, to create and deliver value to targeted customers at a profit. It is mainly grounded on high quality customer related data and enabled by information technology. CRM is an information industry term that helps an enterprise to manage customer relationships in an organized way and helping the company to provide better services to its customers. CRM is the process of managing all aspects of interaction of the company with its existing customers and the new customers, including prospecting, sales and service. CRM applications try to provide insight into and help in improving the company/customer relationship by combining all these views of customer interaction into one picture.

2. Review of Literature

Paresh Tanna et al “Using Apriori with WEKA for Frequent Pattern Mining” Knowledge exploration from the large set of data generated as a result of the various data processing activities due to data mining only. Frequent Pattern Mining is a very important undertaking in data mining. Apriori approach applied to generate frequent item set generally espouse candidate generation and pruning techniques for the satisfaction of the desired objective. This paper shows how the different approaches achieve the objective of frequent mining along with the complexities required to perform the job. This paper demonstrates the use of WEKA tool for association rule mining using Apriori algorithm.

Shrey BavisiÂ et al “A Comparative Study of Different Data Mining Algorithms” Data Mining is used extensively in many sectors today, viz., business, health, security, informatics etc. The successful application of data mining algorithms can be seen in marketing, retail, and other sectors of the industry. The aim of this paper is to present the readers with the various data mining algorithms which have wide applications. This paper focuses on four data mining algorithms K-NN, Naïve Bayes Classifier, Decision tree and C4.5. An attempt has been made to do a comparative study on these four algorithms on the basis of theory, its advantages and disadvantages, and its applications. After studying all these algorithms in detail, we came to a conclusion that the accuracy of these techniques depend on various characteristics such as: type of problem, dataset and performance matrix.

Manjari Anand et al “Customer Relationship Management using Adaptive Resonance Theory” CRM is a kind of implemented model for managing a company’s interactions

Volume 5 Issue 6, June 2016

www.ijsr.net
Licensed Under Creative Commons Attribution CC BY
In this phase, data is acquired using various acquisition tools. This data has to be stored in structured format. Later on mining is performed on the basis of defined rules.

Phase 2
In this phase, structured data has to be normalized and pre-processed. This database will be classified into different classes on rules of GA and Fuzzy KNN classification. After classification various attributes will be extracted from classified data and various parameters are analysed for performance evaluation.

4. Results
In the purposed work different parameters have been used for evaluation of the performance of the purposed work. In this purposed work CRM dataset have been used that has 50000 instances and 231 attributes. This dataset has been used for classification using GAF approach this approach use membership function and genetic operation to classify the data into different classes. To validate the purposed work different algorithm have been used for comparison of the purposed approach. The purposed work has been compared with Fuzzy-KNN classification and KNN classification approach on the basis of different parameters. The dataset have been classified into different classes on the basis of the value of K. the parameters have been computed by using these different values of K.

This table represents the classification of the CRM dataset values when classification has been done on the basis of two classes. These approaches classified the testing data into two different classes on the basis of distance of different samples from the training dataset.

<table>
<thead>
<tr>
<th>Classifiers/metrics</th>
<th>GAF</th>
<th>KNN</th>
<th>FKNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY</td>
<td>0.98</td>
<td>0.87</td>
<td>0.93</td>
</tr>
<tr>
<td>SENSTIVITY</td>
<td>0.92</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>SPECIFICITY</td>
<td>0.16</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>PRECISION</td>
<td>0.92</td>
<td>0.73</td>
<td>0.83</td>
</tr>
<tr>
<td>RECALL</td>
<td>0.95</td>
<td>0.65</td>
<td>0.85</td>
</tr>
<tr>
<td>F-MEASURE</td>
<td>0.97</td>
<td>0.79</td>
<td>0.89</td>
</tr>
<tr>
<td>G-MEAN</td>
<td>0.85</td>
<td>0.73</td>
<td>0.89</td>
</tr>
</tbody>
</table>

This table represents the classification of the CRM dataset values when classification has been done on the basis of three classes. These approaches classified the testing data into two different classes on the basis of distance of different samples from the training dataset. As the number of classes increases in the prediction of the dataset accuracy gets decrease due to availability of actual dataset into two classes.

<table>
<thead>
<tr>
<th>Classifiers/metrics</th>
<th>GAF</th>
<th>KNN</th>
<th>FKNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY</td>
<td>0.76</td>
<td>0.86</td>
<td>0.76</td>
</tr>
<tr>
<td>SENSTIVITY</td>
<td>0.65</td>
<td>0.45</td>
<td>0.55</td>
</tr>
<tr>
<td>SPECIFICITY</td>
<td>0.25</td>
<td>0.55</td>
<td>0.35</td>
</tr>
<tr>
<td>PRECISION</td>
<td>0.65</td>
<td>0.7</td>
<td>0.73</td>
</tr>
<tr>
<td>RECALL</td>
<td>0.79</td>
<td>0.62</td>
<td>0.72</td>
</tr>
<tr>
<td>F-MEASURE</td>
<td>0.87</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>G-MEAN</td>
<td>0.65</td>
<td>0.67</td>
<td>0.71</td>
</tr>
</tbody>
</table>
Table 5.3: Parameters for classification in four classes

<table>
<thead>
<tr>
<th>Classifiers/metrics</th>
<th>GAF</th>
<th>KNN</th>
<th>FKNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY</td>
<td>0.64</td>
<td>0.52</td>
<td>0.6</td>
</tr>
<tr>
<td>SENSTIVITY</td>
<td>0.56</td>
<td>0.52</td>
<td>0.51</td>
</tr>
<tr>
<td>SPECIFICITY</td>
<td>0.56</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>PRECISION</td>
<td>0.65</td>
<td>0.54</td>
<td>0.62</td>
</tr>
<tr>
<td>RECALL</td>
<td>0.59</td>
<td>0.63</td>
<td>0.69</td>
</tr>
<tr>
<td>F-MEASURE</td>
<td>0.69</td>
<td>0.61</td>
<td>0.64</td>
</tr>
<tr>
<td>G-MEAN</td>
<td>0.61</td>
<td>0.55</td>
<td>0.59</td>
</tr>
</tbody>
</table>

This table represents the classification of the CRM dataset values when classification has been done on the basis of four classes. These approaches classified the testing data into two different classes on the basis of distance of different samples from the training dataset. As the number of classes increases in the prediction of the dataset accuracy gets decrease due to availability of actual dataset into two classes.

4.1 Comparison of Evaluation Parameters

Figure 4.1: Comparison of different approaches for K=2

This figure represents the comparison between different classification approaches on the basis of different parameters. These approaches have been used for classification of data into two different classes and the parameters have been analyzed for all the approaches. On the basis of these parameters one can validate and optimized best approach for classification.

Figure 4.2: Comparison of different approaches for K=3

This figure represents the comparison between different classification approaches on the basis of different parameters. These approaches have been used for classification of data into three different classes and the parameters have been analyzed for all the approaches. On the basis of these parameters one can validate and optimized best approach for classification.

Figure 4.3: Comparison of different approaches for K=4

This figure represents the comparison between different classification approaches on the basis of different parameters. These approaches have been used for classification of data into four different classes and the parameters have been analyzed for all the approaches. On the basis of these parameters one can validate and optimized best approach for classification.

5. Conclusion & Future Scope

In the purposed work the fuzzy based membership function has been used for classification this approach assigns the weight age fir dataset attributes on the basis of fuzzy membership rules. After the assignment of weight age the data set distance has been computed using Euclidian classifier and the function has been used that use both weight age and distance factor for prediction of class label to a single dataset. After this classification genetic algorithm has been used for optimization of predicted label. Genetic algorithm use different operators like crossover, mutation, and selection for prediction of best classification labels to a data samples. After this the actual and predicted class labels have been used for prediction of various parameters that have been used for validation of purposed work. The purposed work has been compared with various approaches like KNN and fuzzy KNN.

By analyzing various parameters like accuracy, precision, recall, f-measure, sensitivity, specificity and G-mean one can conclude that the fuzzy and genetic algorithm based approach provide better classification rather than that of the simple fuzzy classification and KNN classification.

6. Future Work

In the future reference the classification can be done by using different parameters in the real world data. In the future the data can be used for classification by using other artificial intelligence approaches for optimization of predicted labels.

References


[2] Nedaabdelhamid, Aladdin Ayesh and FadiThabtah


