Assignment 2: Classification

COMP3308

Table of Contents:

1. Research Aim
2. Data Description
   1. Dataset overview
   2. CFS overview
3. Results and discussion
   1. Classifier performance comparison
   2. Java Classifier implementation vs Weka
   3. Impact of feature selection
4. Conclusion
5. Reflection

310275350 – WSMI6415

450158298 – NZAI0011

Submission Date: 05-18-2018

Research Aim

The research aims of this study sought the following:

* The implementation of two classifiers: Naïve Beyes (NB) and K-Nearest Neighbour (KNN) used with the Pima Diabetes dataset.
* The evaluation of eight classifiers using Weka, a machine learning suite, to gauge their performance.
* The effect of reducing dimensionality via correlation feature selection (CFS) on the Pima Diabetes dataset.

**Why is this problem important?**

A large component of Machine Learning is to choose an appropriate classifier suitable for a dataset. Classifiers are widely used and have many applications across the world. They are used in decision making logic systems such as automated vehicles. In the case of automated vehicles, classifiers are used to make real-time decisions and factors such as the size and dimensionality of the datasets used in these systems impact the performance. The accuracy and speed of these classifier algorithms must be considered carefully in order to minimise the risk of a “bad” decision being made. Hence a solid understanding of the performance of specific classifiers is a valuable topic of research for both academia and private enterprise. Further to this, the usage of a correlation feature selection algorithm is important as it helps improve the performance of a classifier by reducing dimensionality, without significantly reducing the accuracy level.

Data

**Pima Indian Diabetes dataset:**

The data we are evaluating in this research comes from the Pima Indian Diabetes dataset which was originally released in 1990. The Pima Indians are a group of Native Indians residing in the USA. The dataset describes the links between Pima Indian heritage and the incidence of diabetes. The data contains 768 observations of female Pima Indians over the age of 21, containing eight attributes pertaining to their health. The dataset contains the following diagnostic measurements:

**Table 1: Dataset attributes**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Attribute | Unit | Mean | Standard Deviation |
| 1 | Number of times pregnant | Discrete | 3.8 | 3.4 |
| 2 | Plasma glucose concentration |  | 121.7 | 30.4 |
| 3 | Blood pressure | mm Hg | 72.4 | 12.1 |
| 4 | Triceps skin fold thickness | mm | 29.1 | 8.8 |
| 5 | Insulin level | μ U/ml | 155.3 | 85.0 |
| 6 | Body mass index | kg/m2 | 32.5 | 6.9 |
| 7 | Diabetes pedigree function | - | 0.5 | 0.3 |
| 8 | Age | years | 33.2 | 11.8 |
| 9 | Class variable | “yes” or “no" | - | - |

All the attributes except for the class variable are numerical and were normalised before use. Each record in the dataset is classified as either “yes” or “no”. In the dataset, 500 entries yielded a “no” result and 268 yielded a “yes” result.

Correlated-based feature selection (CFS):

We ran the dataset through a CFS algorithm which created a subset of attributes that yielded the biggest change in accuracy. This addresses a key issue known as dimensionality in Machine Learning. There is a trade-off between dimensionality and computational complexity. When the dimensionality is increased, it theoretically should promote accuracy, however this directly increases the complexity within the classifier. Notably, a well-defined dataset contains features which are highly correlated with a class variable, however these features should be independent and therefore uncorrelated with each other.

CFS is an algorithm that seeks to improve the explanatory power of a dataset by pruning attributes and selecting only those attributes which add to the explanatory power of the model. This serves to reduce the dimensionality of the dataset. This is important for the classifiers we are evaluating. In particular, there is a clear benefit of reducing the overall noise within the dataset when running the K-Nearest Neighbour algorithm as it can be impacted by the inclusion of irrelevant variables. These benefits are also evident in the Naïve Bayes algorithm which has an assumption that variables are independent and not correlated, which is not always the case.

Correlation-based feature selection utilises a correlation measure and a heuristic search strategy. Variables are first measured in their correlation to each other and then in their correlation to the class variable. Specifically, the heuristic evaluation function which measures the predictive power of a set of attributes and this is weighed against the amount of correlation between the attributes. In practice, a poor heuristic is yielded if the correlation of a set of attributes with each other is high. Logically, if a set of attributes has a low predictive power, then the heuristic will be poorer.

When we applied CFS within Weka, a tree of attributes is created using the created heuristic. A “best-first search” approach is then applied to find the resulting attributes to be included in the model. It should be noted that the “best first approach” is likely A\* search as it can find the global minima through backtracking.

Table 2: CFS Attributes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Attribute | Unit | Mean | Standard Deviation |
| 2 | Plasma glucose concentration |  | 121.7 | 30.4 |
| 5 | Insulin level | μ U/ml | 155.3 | 85.0 |
| 6 | Body mass index | kg/m2 | 32.5 | 6.9 |
| 7 | Diabetes pedigree function | - | 0.5 | 0.3 |
| 8 | Age | years | 33.2 | 11.8 |

Results and Discussion

The following discussion and results demonstrate the performance of each algorithm ran within Weka and the Java implementations of Naïve Bayes and k-Nearest Neighbours. A 10-fold stratified cross-validation was used and the results in both tables are the mean of each of the folded datasets.

Table 3: Weka accuracy, percentage change for 10-fold stratification across different classifiers.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Weka | ZeroR | 1R | 1NN | 3NN | NB | DT | MLP | SVM |
| Initial Model (%) | 65.104 | 70.833 | 67.838 | 72.656 | 75.130 | 71.875 | 75.391 | 76.302 |
| CFS (%) | 65.104 | 70.833 | 69.010 | 73.307 | 76.302 | 73.307 | 75.781 | 76.693 |
| Change (%) | - | - | +1.172 | +0.651 | +1.172 | +1.432 | +0.390 | +0.391 |
| Ranking | 8 | 6 | 7 | 4 | 3 | 5 | 2 | 1 |
| Ranking  CFS | 8 | 6 | 7 | 4 | 2 | 5 | 3 | 1 |

Table 4: MyClassifier (Java) accuracy for 10-fold stratification

|  |  |  |  |
| --- | --- | --- | --- |
| MyClassifier | My1NN | My3NN | MyNB |
| Initial Model (%) | 74.990 | 69.528 | 73.559 |
| CFS (%) | 76.555 | 68.117 | 71.362 |
| Change (%) | -1.565 | +1.411 | +2.198 |

Performance evaluation:

Accuracy:

Table 3 shows the accuracy before and after running correlation feature selection. Notbaly it shows that the accuracy

Conclusion

Reflection