Enhancing Binary Classification by Modeling Uncertain Boundary in Three-Way Decisions

Abstract:

With the explosive growth of electronic text documents, text classification, one of the crucial technologies of information organization and information filtering, is becoming increasingly important and attracting extensive attention in related research areas in recent years. Text classification plays a key role in both organising and seeking the relevant information (that is usually labeled as "positive") from huge data sets. It is the process of classifying documents into predefined different categories based on their relevance to a topic, a category or a user. There are many practical applications of text classification, such as in news, e-mails, web pages, academic papers, medical records, and customer reviews. A number of text classification techniques have been developed, including support vector machines (SVM), Naive Bayes (NB), Rocchio similarity, k-Nearest Neighbors (k-NN), and decision trees.

Existing System:

Text classification is a process of classifying documents into predefined categories through different classifiers learned from labelled or unlabelled training samples. Many researchers who work on binary text classification attempt to find a more effective way to separate relevant texts from a large data set. However, current text classifiers cannot unambiguously describe the decision boundary between positive and negative objects because of uncertainties caused by text feature selection and the knowledge learning process. This paper proposes a three-way decision model for dealing with the uncertain boundary to improve the binary text classification performance based on the rough set techniques and centroid solution. It aims to understand the uncertain boundary through partitioning the training samples into three regions (the positive, boundary and negative regions) by two main boundary vectors C~P and C~N, created from the labeled positive and negative training subsets, respectively, and further resolve the objects in the boundary region by two derived boundary vectors B~P and B~N, produced according to the structure of the boundary region. It involves an indirect strategy which is composed of two successive steps in the whole classification process: 'two-way to three-way' and ‘three-way to two-way’. Four decision rules are proposed from the training process and applied to the incoming documents for more precise classification.

Proposed System:

This paper proposed an innovative three-way decision approach for addressing the problem of uncertain decision boundary to improve the performance of binary text classification. The experimental results show that the proposed model can significantly improve the performance of the binary text classification in terms of F1 and AUC, and can achieve a high Accuracy compared with other six baseline models. Through this research, the following conclusions can be made. This study has revealed that a satisfactory classifier can be implemented in an indirect way via an intermediate step of three region partitioning, and that the structure and properties of the boundary region obtained at the training stage can be applied to the incoming documents through the two pairs of learned boundary vectors, both of which are based on the theoretical derivation and experimental results.

Modules:

- Three-way decisions.
- Modelling uncertain decision boundary.

SYSTEM REQUIREMENTS

H/W System Configuration:-

Processor	-	Pentium –III
RAM	-	256 MB (min)
Hard Disk	-	20 GB
Key Board	-	Standard Windows Keyboard
Mouse	-	Two or Three Button Mouse

Monitor - SVGA

S/W System Configuration:-

Operating System	:	Windows95/98/2000/XP
Application Server	:	Tomcat5.0/6.X
Front End	:	HTML, Jsp
Scripts	:	JavaScript.
Server side Script	:	Java Server Pages.
Database	:	MySQL 5.0
Database Connectivity	:	JDBC