Introduction
Highly imbalanced data limits the effectiveness of traditional classification algorithms. Many fields such as bioinformatics, national security and information security have highly imbalanced data classification problems[1]. An example of highly imbalanced data classification is finger print recognition where only one print is positive from millions stored in a database.

For some classification problems there may be an extremely accurate classification method. This method can be used as an “Oracle” to correctly classify input data. For finger print classification, presenting the candidate print to a finger print analyst can be considered as using an “Oracle”.

For Starburst landmark detection the Qdoba Ball Region Analysis algorithm gives no known errors, however it is computationally expensive. In order to efficiently process images using this “Oracle” we want to eliminate as many True Negatives as possible before processing the remaining data with the Qdoba algorithm.

Our And Conjuncted Linear Decision Rules (ACLDR) method solves highly imbalanced data classification problems 6000 times faster than the Qdoba ball alone. The ACLDR can quickly eliminate large amount of true negatives by using linear decision rules.

Our work focuses on separating the markers in a image (Figure 1). The data of the image is highly imbalanced: roughly one patch in 100,000 contains a marker. The combination of the ACLDR and the Qdoba leads to a faster approach for classifying our highly unbalanced data.

The comparative performance of our ACLDR vs Support Vector Machines is presented, and in this case, ACLDR has more powerful classification ability and superior efficiency.

Acknowledgments
Thanks to the UWM Electrical Engineering Department for the Teaching Assistantship Support

Approaches
And Conjuncted Linear Decision Rules:

Figure 2. shows the number of data points remaining after using the rules from 1 to 32

1. The ACLDR consists of 32 rules. Each rule evaluates selected features of the input data.
2. Once the point is classified as Negative, it will be eliminated. Figure 2 show points remaining after each rule.
3. The original data has 1028916 True Negative. (Figure 3(a)). 232035 True Negative is cut out by using the first Linear Decision rule (Figure 3(b)). 12009 True Negative is cut out the remaining data from the first cut. (Figure 3(c)).

Support Vector Machines:
Support Vector Machines (SVM) is a standard technique used to perform classification and regression. It is widely used in the areas of Data Mining, Bioinformatics, Image Processing and Artificial Intelligence. SVM is using all support vectors for each point evaluation, and in this test case SVM has 56 times higher error rate and at least 100 times higher time cost.

Results

Figure 4. The Logarithmic Scale Relative False Positive Rate and Relative Run Time from SVM and ACLDR. SVM with adjusted training.

When an oracle is available, it is desired that the first classifier has 100% of True Positive Rate (Sensitivity). In this case, the relative Run Time of SVM classification is from 2 to 3 orders of magnitudes slower than the Run Time used by ACLDR classification and the Relative False Positive Rate is 60 to 85 times higher than that from the ACLDR.

Conclusions and Future Work
Relative to the Support Vector Machines, And Conjuncted Linear Decision Rules is a fast method for imbalanced data classification. It saves at least 100 times computing cost and also reduces by 60x false positives for further classification. These are important characteristics when no False Negative and high efficiency are required.

Currently, we apply the ACLDR in 4 dimensions (with 4 features). Some rules may be able to eliminate more points in higher dimension when the points are not separable in the current dimension. Our future work may attempt to find optimal dimension (number of features) for each rule and find a more efficient features selection technique to select the best combination of features for each rule.

Literature cited