tion by an exponential decay function. Moreover, this allows to reuse node entries if their contribution is too insignificant due to their age. As a consequence, we can maintain an up-to-date view on the data distribution in constant space. The additivity property of the CF [2] allows the comparison of data distributions from arbitrary points in time. Applying a pyramidal time frame as in [1] guarantees a moderate memory consumption even for long running applications.

Through the hierarchical nature of the Bayes tree as an index structure we can insert new objects in logarithmic time and hence can maintain a finer cluster representation than previous approaches in the same time. Moreover, using these fine grained CF representation we can find clusters of arbitrary shape by using density based clustering in an offline component as in [5].

A promising research direction in using index structures for anytime stream mining is the extension of the Bayes tree to enable anytime clustering. This can be achieved by modifying the entry structure such that we can "park" insertion objects in inner nodes and take them along in a later descent. Another great benefit of this modification is the property of *self-adaptation*. More precisely, the size of the tree will automatically adapt itself to the stream speed since insertion objects will descent as far as time permits, be parked there and hence no further splits occur.

Further topics include the detection of outliers, handling of missing values and the investigation of a subspace variant of the Bayes tree. In general we believe that the effective usage of index structures for stream data mining as described in [16, 13] and Section 3 can be successfully extended.

5. CONCLUSION

We have presented and applied a novel index-based anytime classifier (Bayes tree) in recent work constituting our first goal in enabling anytime Bayesian classification. Our next goal is the improvement of its performance. To this end we presented ongoing work in this paper wherein we improved the anytime classification accuracy up to 13% through different bulk loading approaches. Moreover, we have laid out various further research aspects regarding index-based stream classification. The third goal is the extension of our technique to other data mining tasks. Hereto we identified approaches to use index structures for modeling evolving data streams or even for anytime clustering.

Acknowledgments

This work has been supported by the UMIC Research Centre, RWTH Aachen University.

6. REFERENCES

[1] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In *29th VLDB*, 2003.

[2] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for projected clustering of high dimensional data streams. In *30th VLDB*, 2004.

[3] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. On demand classification of data streams. In *10th ACM KDD*, pages 503–508, 2004.

[4] B. Arai, G. Das, D. Gunopulos, and N. Koudas. Anytime measures for top-k algorithms. In *33rd VLDB*, 2007.

[5] F. Cao, M. Ester, W. Qian, and A. Zhou. Density-based clustering over an evolving data stream with noise. In *SDM*, 2006.

[6] J.-Y. Chen, J. Hershey, P. Olsen, and E. Yashchin. Accelerated monte carlo for kullback-leibler divergence between gaussian mixture models. In *ICASSP*, 2008.

[7] D. DeCoste. Anytime interval-valued outputs for kernel machines: Fast support vector machine classification via distance geometry. In *ICML*, 2002.

[8] A. P. Dempster, N. M. L. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[9] S. Esmeir and S. Markovitch. Anytime induction of decision trees: An iterative improvement approach. In *21st AAAI*, 2006.

[10] J. Goldberger and S. T. Roweis. Hierarchical clustering of a mixture model. In *NIPS*, 2004.

[11] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *SIGMOD*, pages 47–57, 1984.

[12] S. Hettich and S. Bay. The UCI KDD archive http://kdd.ics.uci.edu, 1999.

[13] P. Kranen, D. Kensche, S. Kim, N. Zimmermann, E. Müller, C. Quix, X. Li, T. Gries, T. Seidl, M. Jarke, and S. Leonhardt. Mobile mining and information management in healthnet scenarios. In *9th IEEE MDM*, 2008.

[14] S. T. Leutenegger, J. M. Edgington, and M. A. Lopez. Str: A simple and efficient algorithm for r-tree packing. In *ICDE*, pages 497–506, 1997.

[15] G. S. Manku and R. Motwani. Approximate frequency counts over data streams. In *28th VLDB*, 2002.

[16] T. Seidl, I. Assent, P. Kranen, R. Krieger, and J. Herrmann. Indexing density models for incremental learning and anytime classification on data streams. In *12th EDBT/ICDT*, 2009.

[17] A. Silberstein, A. Gelfand, K. Munagala, G. Puggioni, and J. Yang. Suppressions and failures in sensor data: A bayesian approach. In *33rd VLDB*, 2007.

[18] B. Silverman. Density Estimation for Statistics and Data Analysis. 1986.

[19] P. Stone and D. Andre. Physiological data modeling contest (ICML-2004): http://www.cs.utexas.edu/users/pstone/workshops /2004icml/, 2004.

[20] K. Ueno, X. Xi, E. J. Keogh, and D.-J. Lee. Anytime classification using the nearest neighbor algorithm with applications to stream mining. In *ICDM*, 2006.

[21] N. Vasconcelos and A. Lippman. Learning mixture hierarchies. In *NIPS*, pages 606–612, 1998.

[22] M. Vlachos, J. Lin, E. J. Keogh, and D. Gunopulos. A wavelet-based anytime algorithm for k-means clustering of time series. In *Workshop on Clustering High Dimensionality Data and Its Applications*, 2003.

[23] H. Wang, W. Fan, P. S. Yu, and J. Han. Mining concept-drifting data streams using ensemble classifiers. In *9th ACM KDD*, 2003.

[24] Y. Yang, G. I. Webb, K. B. Korb, and K. M. Ting. Classifying under computational resource constraints: anytime classification using probabilistic estimators. *Machine Learning*, 69(1), 2007.